VISIT…

LANZAROTE
Caliente.COM

二〇一四 第一卷 第一期
语料库语言学
外研社

语料库语言学

CORPUS LINGUISTICS

1
Vol.1 No.1
第1卷 第1期
2014

北京外国语大学中国外语教育研究中心

Annotation
AntConc
BNC
Chunk
CLEC
Cluster
Collocation
Concgram
Concordance
Context
co-occurrence
Corpora
Corpus
corpus-based
corpus-driven
Frequency
lexical item
KWIC
Wordlist
Keywords
Chunk
Node
Text
Lemma
POS
Tagging
Lexis
N-gram
Annotation
Phraseology
Idiom principle
Open Choice principle
Pattern
Units of meaning
Semantic preference
Semantic prosody
Sinclair
Data driven learning
Meta-data
Regex
WordSmith
CLEC
SWECCL



外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

语料库语言学
（半年刊）

Corpus Linguistics
(Biannual)

主　管：中华人民共和国教育部
主　办：北京外国语大学
承　办：中国外语教育研究中心
出　版：外语教学与研究出版社

Administered by the Ministry of Education of China
Directed by Beijing Foreign Studies University
Edited at the National Research Centre for Foreign Language Education
Published by Foreign Language Teaching and Research Press

主　编：梁茂成、许家金
编　校：吉　洁、刘　燕

Editors: Liang Maocheng and Xu Jiajin
Proofreaders: Ji Jie and Liu Yan

编审委员会（按姓氏音序）
冯志伟（杭州师范大学）
顾曰国（中国社会科学院）
桂诗春（广东外语外贸大学）
何安平（华南师范大学）
胡开宝（上海交通大学）
李文中（北京外国语大学）
刘泽权（燕山大学）
陆小飞（美国宾州州立大学）
濮建忠（解放军外国语学院）
陶红印（美国加州大学洛杉矶分校）
王克非（北京外国语大学）
卫乃兴（北京航空航天大学）
文秋芳（北京外国语大学）
肖忠华（浙江大学）
杨惠中（上海交通大学）

Editorial Board (in alphabetical order)
Feng Zhiwei (Hangzhou Normal University)
Gu Yueguo (Chinese Academy of Social Sciences)
Gui Shichun (Guangdong University of Foreign Studies)
He Anping (South China Normal University)
Hu Kaibao (Shanghai Jiao Tong University)
Li Wenzhong (Beijing Foreign Studies University)
Liu Zequan (Yanshan University)
Lu Xiaofei (The Pennsylvania State University)
Pu Jianzhong (The PLA University of Foreign Languages)
Tao Hongyin (University of California, Los Angeles)
Wang Kefei (Beijing Foreign Studies University)
Wei Naixing (Beihang University)
Wen Qiufang (Beijing Foreign Studies University)
Xiao Zhonghua (Zhejiang University)
Yang Huizhong (Shanghai Jiao Tong University)

电　　话：（010）88816828
电子邮箱：bfsucrg@sina.com

本刊地址：北京市西三环北路19号北京外国语大学中国外语教育研究中心
《语料库语言学》编辑部（100089）

《语料库语言学》
2014年　第1卷　第1期

目　录

发 刊 词

创刊寄语

学者聚焦

语料库语言学答客问 .... 桂诗春（1）

研究论文

英语搭配框架的意义单位再探 .... 何安平（16）
语料库、平义原则和美国法律中的诉讼证据 .... 梁茂成（25）
四大名著汉英平行语料库的宏观语言特征研究 .... 刘泽权、刘鼎甲（34）
中国学者应用语言学英语论文中的词块研究 .... 郑红红（47）

研究综述

语义韵研究的理论、方法与应用 .... 陆　军（58）
基于CiteSpace的国内语料库语言学
　　研究概述（1998–2013）.. 刘　霞、许家金、刘　磊（69）

研制开发

人文社科学术文本俄汉平行语料库的创建与研究 .... 陶　源（78）

史料重刊

语体文应用字汇……陈鹤琴（94）

会讯动态

第36届ICAME大会征文通知……（103）
“国际学习者语料库研究会”成立……（103）
“ToRCH2009现代汉语语料库”建成……（103）

《语料库语言学》稿约……（105）
本刊格式体例……（106）

英文摘要……（110）

CORPUS LINGUISTICS

Volume 1, Number 1, 2014

Table of Contents

Inaugural editorial

Congratulations from scholars home and abroad

Corpus linguists in perspective

Some reflections on Corpus Linguistics upon request........................... *GUI Shichun* (1)

Research articles

Revisiting English collocational frameworks as units of meaning......... *HE Anping* (16)

Corpora, the Plain Meaning Rule and litigation proofs in the US law... *LIANG Maocheng* (25)

A study of the macro linguistic features of the English translations of the Four Chinese Classic Novels .... *LIU Zequan & LIU Dingjia* (34)

A study of the use of chunks in English research articles by Chinese EFL scholars .... *ZHENG Honghong* (47)

State-of-the-art reviews

Research on semantic prosody: Theory, methodology and application........ *LU Jun* (58)

An overview of corpus linguistics research in China (1998–2013): A CiteSpace based analysis.............................. *LIU Xia, XU Jiajin & LIU Lei* (69)

New corpora, tools and methods

Construction of and research on parallel Russian-Chinese corpora: With special reference to academic texts of social sciences and humanities.. *TAO Yuan* (78)

Historical text reprinted

Characters used in vernacular Chinese .... *CHEN Heqin* (94)

发刊词

语料库，顾名思义，即语言材料数据库。其对应的英文corpus源自拉丁文，有“身体”、“尸体”、“~体”、“汇集”等义。corpus的本义，除见于一些专业术语和固定表达外，已很罕用。与现代语言学概念中corpus相近的用法则有a corpus of Shakespeare（莎翁作品集萃）。如今，语言学中的corpus，则专指服务于语言研究及应用，并存储于计算机中的大规模真实语言材料的文本汇集。

现代大型电子语料库及相关研究，始于20世纪60年代大西洋两岸，经50余年发展，至今几成显学。语料库语言学是一门朝阳学科：注重语言事实的量化描述，分析流程规范，逐渐赢得学界认可。然而，有观点认为，语料库方法强于语言用法特征的描述，弱于理论阐释。因此，若止于量化，终于描述，语料库语言学这株苗圃新花，将注定昙花一现，难以长盛。语料库语言学除去坚守自身理论内核外，还具有包纳其他语言学理论的开放性。有必要向其他语言学流派，或其他学术领域虚心求教，以求学科臻于完善。

在语料库研究成果应用方面，西方学者已迈出坚实步伐，在语法书编写、词典编纂、教育教学上都有诸多经典案例。国内语言学界利用语料库者也已不在少数，对相关理念、方法的把握并不落后于西方。我们期待具有重大影响力的学术成果问世，这自然需要学界同仁不断精进。

在我国，语料库语言学从萌芽到蓬勃发展，已有30多年。30多年来，中国的语料库语言学者筚路蓝缕，从草创到开拓，从试验到创新，功不可没。在他们身影的背后，留下一行行清晰的脚印。他们创建语料库，培育团队，立足本土开展语料库应用，引介与创新并重，如今已成果缤纷，蔚然可观。在今天的中国语料库研究大“家庭”里，汇聚了德高望重、笔耕不辍的长者，年富力强的中坚，更有一批精通计算机技术的专业人士与我们同行，共同构建了一个富有活力的学科队伍。

编辑《语料库语言学》，意在记录和追踪我国语料库语言学研究的进展和动态，使之成为了解语料库语言学研究的一个窗口。我们以此期望，中国语料库语言学界能作出既具本土特色，又能与国际对话的优质研究。

《语料库语言学》每年刊出两期。编辑工作主要由北京外国语大学中国外语教育研究中心语料库语言学团队完成。在刊物筹办和编校过程中得到国内外专家和外语教学与研究出版社的鼎力支持，在此致以最诚挚的谢意。

创刊寄语

This new journal, *Corpus Linguistics*, edited at the National Research Center for Foreign Language Education at Beijing Foreign Studies University, shows strikingly that corpus linguistics has now become a truly global enterprise, with a multitude of approaches competing for new ways to look at language from varied perspectives. These days, Chinese corpus studies display cutting edge research and are keenly reviewed all over the world. *Corpus Linguistics* will promote new directions in our discipline, thus making a crucial contribution to language studies in general.

Wolfgang Teubert, Birmingham University, UK

《语料库语言学》的创刊是我国语言学界一件值得庆贺的大事。在我国推进以方块字为基础的汉语语料库研究面临着许多研究课题，也需要一个平台来交换怎样对方块字进行信息加工的理论和研究成果，以集思广益。在外语教学中利用语料库方法也是一个公认的很好的手段，但是目前也只限于英语，还有许多外语，如俄语、法语、德语、西班牙语等等，也有不少学习者，也需要推广语料库方法。衷心祝愿这个刊物能够贴近我国实际，办出其特色。

广东外语外贸大学　桂诗春

《语料库语言学》是我国第一本语料库语言学的专门刊物，热烈祝贺《语料库语言学》创刊！语料库语言学为语言研究的现代化提供了强有力的手段，使得语言研究不仅依靠语言研究者明察秋毫的洞察力，而且主要依靠对于大规模语料中复杂纷繁语言现象的细心观察，从而尽可能地避免了主观性和片面性，这对于语言学的现代化具有不可估量的作用。

杭州师范大学　冯志伟

热烈祝贺《语料库语言学》杂志创刊！

语料库语言学以真实语言数据为研究对象，凭借计算技术，采用数据驱动的实证主义研究方法，从宏观的角度对大量语言事实、对语言交际和语言学习的行为规律进行多层面的研究，尤其是有关语言使用的概率信息，为语言学研究提供了新的途径、带来新的理念和方法。这方面的研究必然有助于人们加深对语言本质的理解，研究成果可以直接应用于语言教学。正因为如此，语料库语言学研究队伍在我国近年来不断壮大，现在终于有了一本自己的专业刊物，为语料库语言学研究工作者开展学术讨论、发表研究成果、交流学术经验提供了园地，希望我国年轻一代学者不断开拓创新，在语料库语言学理论研究、工具开发、成果应

用方面不断取得丰硕成果，为世界语料库语言学的发展作出我们应有的贡献。

上海交通大学　杨惠中

从学术发展的观点看，语料库语言学方兴未艾，它便于我们更仔细和更广范围地观察和认识以往因条件所限未能认识的语言现象，开辟出一个讨论这类问题的学术园地，是令人鼓舞的。

从学术发表的角度看，目前中国（包括港澳台地区）尚无一家专门的语料库语言学刊物，大量存在的是综合性刊物，这不利于学术的专精和深入。国际上这一专业刊物也不多见，故可以吸收国外学者论述，形成国际化学术平台。

从学刊编辑的成员看，即将问世的这本刊物，其编辑都是语料库语言学研究的专家和新秀，这对于识稿、组稿、审稿、改稿以及作者、读者、编者的互动都非常有益、非常重要，可知《语料库语言学》前景一片光明。

北京外国语大学　王克非

喜悉《语料库语言学》付梓问世，倍感欣慰。坚信必将成为学人案头不可或缺之物。

中国社会科学院　顾曰国

求真，存真，至真。

北京航空航天大学　卫乃兴

语料库语言学答客问

广东外语外贸大学 桂诗春

编者按

本期“学者聚焦”关注的是桂诗春教授。桂教授是我国外语界语料库语言学研究的先行者之一。他同杨惠中教授主持创建的“中国英语学习者语料库”，极大地促进了我国英语中介语的实证研究。桂先生年过耄耋，仍然紧跟语料库研究最新技术和方法。他79岁高龄时出版了基于自建学术英语语料库的多维度英语语体研究专著。近期，他还自学R语言，以用于英汉语语料的统计分析。

为能让更多年轻后学从桂先生身上汲取学术养分。本刊特于创刊号登载对桂先生的专访，以飨读者。

1. 您最早是什么时候开始接触语料库的？您能描述一下当时国内语料库研究开展的情况吗？

世界上第一个机读英语语料库（布朗语料库，Brown Corpus）建于20世纪60年代中叶。当时我国正值“文革”，与国外隔绝，直到“文革”结束后，我才开始接触语料库语言学。首先看到的是Kučera和Francis的*Computational Analysis of Present-Day American English*，那是100万词次的布朗语料库的文字描述版，不久又看到John Carroll等人基于500万词次的*Word Frequency Book*，虽然两者都不是直接可用的电子化语料库。但最早引起我兴趣的是心理语言学家John Carroll为这两本语料库所写的《序言》，然后又看到Gustav Herdan所写的两本书：*Type-Token Mathematics*（1960）和*Quantitative Linguistics*（1964）。当时还没有语料库语言学的提法，但这两本书和Carroll的《序言》，却给我打下了语料库的理论和数学基础，开始认识到通过语料库调查进行语言研究的重要意义。上海交通大学杨惠中、黄人杰等人的团队，也差不多在这一阶段认识到语料库的前景，并开始在我国建立自己的语料库；他们收集并创建了JDEST（Jiao Da English for Science and Technology）语料库，并基于该语料库来编制科技英语常用词表。其间我也访问过他们，并在现场看过他们的成果。但是布朗语料库也好，JDEST语料库也好，当时都是依托大型计算机来完成的。而我所在的单位并没有计算机，于是就向上级申请购买一台Apple II型的微型计算机。教育部门领导最初的反应是：你们又不是工科院系，要什么计算机？经过我们努力说明和争取，最后购进了3台，分给几个部属外语学院（北外、上外和广外）。

当时的计算机技术远没有现在发达，中央处理器和内存都较低级，外部储存手段只有5英寸软盘，光学扫描仪还没有问世。1985年，我招了一个硕士生祝启波，他原在石油大学广州分院教英语，也上过计算机课，于是我们就开始在一个IBM PC/XT计算机平台上，开发石油英语语料库GPEC（Guangzhou Petroleum English Corpus）。祝走访了我国石油系统的几个院系，根据石油探测、石油提炼和石油探钻三大类进行采样和人工输入文本，而

且在一台微机上，进行文件的组合、整理和运算，终于建立了一个40万词次的石油英语语料库。这个语料库最后以《石油英语频率词典》（1991）的名义发表，使用的是Carroll的*Word Frequency Book*的几个统计量（U、SFI、D和F）。我在为该书所写的《序言》里不得不说：The build-up of corpora requires a Brobdingnagian effort，（Brobdingnag是《格里佛游记》里的“大人国”），这个研究的成果不但是一个石油英语语料库，而且还建立了一个在多数人都能拥有的廉价计算机上建立专门用途语料库的模型。Leech（1997：18）在回顾“专门用途语料库”时说过，“这些语料库通过不同手段在逐步增加，首先是敏锐的专门用途语言学家和教师开发自己的语料库，早期的例子是JDEST和GPEC，两者都来自中国。”Leech所不知道的是GPEC是在技术条件那么差的情况下完成的。

至于和语料库有关的软件，最早接触到的是加拿大多伦多大学Ian Lancashire等人开发的TACT2.1，那是在DOS3.0基础上开发的，具有很多英语文本（主要为文学文本），当年可从该大学网站下载使用。TACT已经具有语料库的各种功能（检索、词频表等），不过它的界面并不十分友好。另一个是WordCruncher，主要是一个检索工具，其好处是可以检索汉语，但不能对汉语进行分词。Mike Scott的WordSmith Tools的各个版本都在Windows的环境下运行，把各种功能都组合在一起，且提供不少统计数据，应是一个突破。还应提出的是ICAME在1999年发行了一张光盘，叫做*ICAME Collection of English Language Corpora (2nd Edition)*，其中包括了6个软件（除前述3个外，还有Lexa、Lingfont、Qwick）和20个语料库，规模达1千7百万词次。这张光盘对普及和推进语料库研究，起了很大作用。

2. 那么语料库语言学在国外的发展又如何呢？

布朗语料库问世后，并未引起美国语言学家的注意，因为当时正是生成语言学当道，但在欧洲却起了重要的催生作用。1977年在挪威成立了ICAME（International Computer Archive of Modern and Medieval English）协会，对英语语料库的推广起了重要作用。Simpson & Swales（2001）不得不承认语料库语言学在最近15年的很多发展都来自欧洲，特别是英国和北欧等国学者的研究。其原因是复杂而又有趣的：首先是在北美，理论语言学，因为受到Chomsky的影响，把注意力指向语言结构，即所谓I-language（内部语言），而不是语言使用；其次是在欧洲，特别是对北欧语言学家来说，语言学主要强调语言和社会生活的联系，这是英国语言学家Firth所建立的传统，他提出的“行动中的语言”（Language in action）和“作为使用的意义”（Meaning as use）是这一传统的两个孪生口号（见Leech 1974：71）。

其实“语料库语言学”的说法，是在20世纪八九十年代兴起的，一般把布朗语料库（1967）的发表作为一条分界线，分为前计算机和后计算机（机读）两大阶段：前计算机阶段通常被称为计量语言学（Quantitative Linguistics）、统计语言学（Statistical Linguistics）、机械语言学（mechanolinguistics）等等，Herdan（1966）曾经把这个时候的语言研究归纳成“作为机遇和选择的高级语言理论”：统计语言学就是把语言作为机遇（Chance），而文体统计学（Stylostatistics）则把语言作为选择（Choice）。计量和统计的核心是频数，例如圣经索引（在我国，对一些经典著作都编有Index，被称为“引得”）、词典和常用词表编制、语法和用法调查等等。其中最受人注目的是Quirk等人所作的“英语用法调查”（Survey of English Usage）。根据Svartvik（2007）的回忆，他在1961年就参

与这项研究，当时还没有用corpus这个词，Quirk最初想用descriptive register（描写性语体）、primary material（基本材料）、texts（文本）这几种提法，连corpus的复数是corpuses还是corpora，还拿不定主意，最后有人说，“我想应该是corpi”。Svartvik还记得1963年W. Nelson Francis从布朗大学带来一大堆计算磁带造访Quirk在伦敦大学学院的办公室，这就是他们刚刚完成的机读语料库，标有habeas corpus（拉丁语：意为“人身保护令”，所以corpus实为body（本体）[1]，在英语用法调查基础上，Quirk等人先后编了两部现代英语语法：《现代英语语法》（1972）和《英语语法大全》（1985）。具有同样意义的是Edward Thorndike从1921年到1944年所编制的《教师词汇手册》，把语料规模从10,000词增加到30,000词并按词频排列，所依据的语料规模达450万词。均是在没有计算机支持下完成的。他所编制的*Thorndike Junior Dictionary of English*对常用3,000词作了标记。用手工来排列词频，十分繁复。再如在早期，大主教Hugh动用了500名僧侣来进行拉丁语圣经索引的编纂，后来Alexander Cruden以惊人毅力用两年来完成，但他每天要工作18小时。

布朗语料库开启了后计算机时代，由于欧洲语言学家起了“接棒”的作用，1983年在荷兰Nijmegen召开了一次ICAME会议，主题是“语料库语言学：计算机语料库在英语研究中的使用”，由此语言库语言学的说法就说开了。但Jan Aarts则指出，他在1980年就开始使用荷兰语corpustaalkunde（相当于英语“语料库语言学”）。在70年代以后，机读语料库随着计算机技术（如网络、中央处理器、内存、外部存贮手段、光学阅读器）的开发和发展有了迅猛发展。Renouf（2007）分60、80、90、98、05年代等5个阶段描述了机读语料库如何从100万词发展到几千万和上10亿词，一直到把整个网络作为语料库，因而出现GRID的说法（原意为输电网的线路网，或称为“栅极”，即用户在需要用电就把插头插到插座里，无需知道电源在哪里。）这是把网络作为语料库的结果，因为网络资源爆炸，需要很多索引来使用语料本身，这些索引甚至比语料本身还要多，需要开发软件来把它们组织和存储在“网间数据栅”，这个新系统需要更多的内容标注，这就是计算语言学家所致力设计的“语义网”（semantic web）。

3. 您刚刚提到“生成语言学当道”，这是不是意味着语料库和生成语言学是不相容的呢？

确实，布朗语料库产生后，就受到Chomsky的批判。Chomsky反对的是结构主义和行为主义。早在20世纪50年代Fries在《英语结构》（1952）里使用过会话语料对英语结构进行分析，Chomsky在反对结构主义过程中出版专著《句法结构》（1957），他从一开始就反对根据语料来决定语言的语法性。其实这接触到现代语言学的一个根本问题，Saussure的“语言”（langue）和“言语”（parole），在Chomsky语言学里就是“语言能力”（linguistic competence）和“语言运用”（linguistic performance），后又改称为I-language 和E-language（内部语言和外部语言）。Chomsky虽然也承认这两者的区别，但认为语言学研究的中心应该是语言能力。这就形成语言学研究的两大流派：一派是生成语言学，其哲学基础是理性主义；另一派是功能语言学（Firth、Halliday等）。和功能语言学站在一起的不但有语料库语言学，还有语用学、历史语言学、认知语言学、社会语言学等等，其哲学基础是经验主义。如果站得更高一点来看，前一派关心的是语言中What is possible?（“哪些是可能的？”，即语言能力所容许发生的），而后一派关心的是语言中What is probable?（“哪

些是极有可能的？”，即在语言运用中被使用的概率有多大？）。例如Chomsky所举的著名例子：Colorless green ideas sleep furiously（“无色的绿思想疯狂地睡觉”），在生成语言学者看来，这样的句子是possible（可能的），因为它完全符合英语语法。而Furiously sleep ideas green colorless则是impossible（不可能的），因为不符合英语语法。在语料库学者看来，一般人（除了生成语言学家的专门论述外）是没有什么可能说这样的两句话的，所以那是improbable（极不可能）的。Possible 和probable在英汉词典里都有“可能”的意思，但是前者感兴趣的是有无可能，这是两分法的；而后者则和概率行为有关，是有梯度的。所以“语言”和“言语”其实是一个硬币的两个方面，它们是互补，而不是对立的。两大语言学传统其实是从不同角度来观察语言事实，Halliday（1991）把它们比喻成climate（气候）和weather（天气）以示区别。Newmeyer（2005）是一位生成语言学者，他从语言类型学的生成主义视角来考察possible（实际上是biologically possible“生物学的可能”）和probable的语言，专门讨论了生成主义和功能主义：功能主义学派对其可能也感兴趣，但它认为“语言理论的主要目标是把极有可能和可能区分开来”（Most adherents of the functional school see it [Universal Grammar] as a major goal of linguistic theory to distinguish the possible from the probable）。Newmeyer虽然坚守生成语言学的立场，认为“把语法元素和概率联系起来的证据十分薄弱”，但却指出“功能主义的解释和形式生成语法是完全相容的”，认知语言学家Langacker所提出以用法为基础的语法模型也不赞成在语言知识和语言使用之间作严格的区分。Newmeyer在书中多处用了以频数为基础的解释，来说明什么东西使语言有更大可能（probable），而使语言有可能的（possible），则是Chomsky的普遍语法。Dryer（2007）在对Newmeyer的书评里说，“我是一个类型学家和功能主义者，但我认同Newmeyer多数说法。”McEnery & Wilson（2001）关于Chomsky和语料库的关系也有过详尽论述，认为他对早期语料库语言学的批评（如过于偏态）不无好处，这反而使后来语料库的采样具有更大代表性。

4. 语料库语言学研究的哪些特点最吸引您？

英语对我来说始终是一门外语，即算是按生成语言学的说法，我所具有的语言能力也是汉语的语言能力，自问对一门外语的了解和掌握无法和母语使用者相比。所以使用英语时，觉得没有多大把握时就要向母语使用者请教，但是母语使用者也有其年龄、时代、文化、接触面等等局限。最好的办法是查大型语料库，甚至Google，如果都没有人这样用，就要十分小心。另外通过不同语料库的频数比较，也可以发现许多语体（包括我国英语学习者的英语）的特点。频数的分布可以帮助人们更准确地理解哪些词使用得最多，这对制定常用词表大有益处。我是教英语的，常对其语法变化和发展感兴趣，正如Keller（1994）所指出的，这是间乎“自然”与“人工”之间的第三种现象，可称为“无形之手”（The Invisible Hand），语言和交际就等于市场、贸易、货币一样，它们不是任何人类设计的产物，而是人类活动的结果。就等于“花园小径”一样，它虽然是人走出来的，但却不是具体的哪个人在哪个时候走出来的，而是有人先那么走，别人也觉得这样走比较方便，慢慢也顺着走，走多了就成为“小径”了。语料库的方法更容易昭示这些规约性结果。利用这些结果来编撰语法和词典，这与历史主义的原则更为一致（如Jespersen、Quirk、Biber

编制的英语语法和OED、Collins COBUILD、Longman等词典所收集的例句……等），因为对我们那些把英语作为外语的人来说，实在无法运用自己与生俱来的“语言能力”（像Chomsky所说的，如果有，也只指自己的母语）来进行判断。例如在英语口语里，像Did you want more coffee?这样的句子和过去时无关，而是一个有礼貌的请求，对句子的回应是No，I’m fine（现在时）或Yeah, I’ll have one（将来时）（见Conrad & Biber 2009）。像这样的语言能力对把英语作为外语的学习者来说，只能在特定的语言环境通过接触而学到，而不是生而知之。

5. 有没有哪（个）些学者或某（个）些论著在语料库研究方面对您影响较大？如有的话，您能说说影响主要体现在什么方面吗？

任何一门学科的发展都依赖于这个学科参与者的共同努力；他们在各个方面都作出了自己贡献，不可忽略。总体而言，语料库语言学并非我唯一的学术兴趣，我最早的兴趣是在中国引进和发展应用语言学，后来是心理语言学和语料库语言学，最近又转向语言的进化和演变。最早吸引我的是语料库的研究手段，觉得它和计算机科学结合起来，可以省去很多精力，具有无限广阔前景。语料库语言学之所以有今天的发展，有赖于这个学科建设者各方面不懈努力，在英国有几个中心，包括以Quirk为首的伦敦大学学院（University College London），以Leech为首的兰卡斯特大学（Lancaster University），以Sinclair为首的伯明翰大学（Birmingham University），他们都孜孜不倦地开发和利用语料库，硕果累累，而它们所培育的力量在欧洲各个国家如瑞典、丹麦、意大利、荷兰、德国、比利时等地开花结果。至于我自己并没有从一开始就把语料库语言学作为自己的专业方向，虽然收集了不少论述，也没有一一通读，只是选读其中一些，觉得很有收获，但也不足为训。我最早读到的是Quirk（1968）的*The Use of English*, 因为这本书一出版后，国内就有影印版。书写得深入浅出，饶有兴趣，可能就是他们英语用法调查的副产品。跟着就是Greenbaum & Quirk（1970）对他们调查所使用方法的介绍。至于根据调查所编写的几本语法均属于参照性语法，虽然常参阅，但没有通读。Leech在兰卡斯特大学成立UCREL（Unit for Computer Research on the English language）后不但建立了LOB，而且在Garside，Leech & Sampson（1987）里，提出了词类标记（POS tagging）系统。Sampson还针对Chomsky的生成语言学出版了*Educating Eve*（1997）和*Empirical Linguistics*（2001），并与McCarthy编辑了一部收录语料库重要文献的读物（2004）。Sinclair的*Corpus, Concordance, Collocation*（1991）、*Reading Concordances*（2003）和*Trust the Text*（2004）强调一切以文本为依归，语篇分析和语料库是语言研究的两大支柱。它们的结合有两点好处：1）我们可以对文本提出很多假设，然后用计算机的手段来加以证实。2）它们所处理的型式维度都比语言学习惯于处理的要多一些。Sinclair因此提出考察语篇的必要性，并针对“自由选择原则”（open choice principle）提出“习语原则”（idiom principle），由此开拓了许多检索和搭配的研究。他所领导的团队不但开发了4.5亿词的Bank of English，并据此编制Collins COBUILD高级英语学习者词典和一套包括习语、语法、构词法、动词短语、商业英语、科技英语在内的丛书。他和Renouf（1988）又提出词汇语法和词汇大纲，并指导编写初级英语教程（Willis 2009）。虽然美国受到Chomsky的影响，但是也有一些学校和语言学家坚守这个阵

地，如美国密歇根大学的MICASE（Michigan Corpus of American Spoken English），Biber *et al.*（1999）主持的“朗文口语和书面语语法”，对语体的研究（1988）。Mark Davies在杨百翰大学（Brigham Young University）创建了一个在线语料库平台（http://corpus.byu.edu）可以检索现代美国英语（COCA，4.5亿词）、历史美国英语（COHA，4.5亿词）、全球（20个国家）以网络为基础的英语（GloWbe，19亿词）、英国国家语料库（BYU-BNC，1亿词）、加拿大英语（Strathy，5,000万词）、《时代周刊》语料库（Time Magazine Corpus，1亿词）、美国电视剧语料库（Corpus of American Soap Operas，1亿词），可谓蔚为大观。

6. 您如何评价中国语料库研究在过去若干年的发展以及目前的现状？

前面谈过，中国语料库研究虽然起步较晚，但发展非常迅速，而且很快就变成一门热门学科，成为研究生首选研究方向之一。原因也很简单，因为各种语料库很多，研究工具很普及，解决了资源和方法论的问题；但从学科的发展来看，却隐藏着一些危机，首先是作为一门交叉学科，选择这个方向的研究生应该掌握哪些基本知识？如果对这些知识一无所知或一知半解，则研究很难有什么深度，更难说有什么创新。所以要解决学科定位和学科建设的问题。我觉得语料库语言学是靠几个支撑学科发展起来的，所以需要掌握几门核心课程，如：1）“普通语言学”（包括语言理论、语音、语法、词汇、语义和语用等），它是统揽全局、不可或缺的基础知识；2）“语料库语言学的理论、发展和方法”，这当然是这个专业方向的核心课程；3）“计量（统计）语言学”，这是语料库的基本方法论，但却需要一些数学和统计学的基本知识。我有点怀疑我国有多少语料库语言学研究者是认真读过Oakes（1998）和Manning & Schutze（1999）的，所谓“认真”不仅是指读通，而是亲自动手做过运算的，起码是了解其计算流程的。即使是使用WordSmith Tools等工具，也需要充分利用其各种功能。所以计算机编程能力也是不可少的；4）“文本分析”，或称语篇分析，或批判性语篇分析。如果说语料库制作软件是研究手段的话，那么文本（包括习语、型式、口语与书面语、语域、文体、专门用途语言等等）就是其研究对象，文本分析在计算机支持下得到很大发展，见Stubbs（1996），Carter（1997），Adolphs（2006），Baker（2006）。上述四个方面的知识，都是以语料库语言学作为研究对象的学者所必须具备的，也是建立语料库语言学这个学科都应该开设的核心课程。在语料库语言学日益兴旺的今天，那些热切地希望建立这个学科的单位都必须考虑培养、引进这些方面的精英与才智，不然的话就会出现最初是“一哄而起”，然后是人才断层的问题。语料库语言学在我国的路子就会越走越窄，目前我所看到的一些研究，较多的限于一些频数的罗列和比较；有些研究也使用到一些多维度的研究手段如因子分析，但研究者是依靠SPSS算出来的，而且并没有用在点子上，一些主要数据（如因子负荷）并没有列出和解释（对分离出的几个因子提出假设是因子分析的主要目标）。Gries（2009，2013）写过两本关于怎样使用R语言来处理语料库和语言学中的描写统计学与分析统计学，不但介绍了它们的基本原理，而且引导读者用R来编制程序。这两本书都值得学习和亲自动手运作，它有助于我们摆脱对现行商业程序的依赖，真正了解内部机理。

7. 您能谈谈中国语料库研究在国际语料库研究学界应如何自我定位？（比如在选题、理论视角、方法论等方面）

我国具有悠久文明历史，典籍浩繁，我觉得中国语料库语言学应该首先定位在对汉语的研究；那是我们的母语，责无旁贷。西方语料库是在处理拼音文字基础上发展起来的，怎样处理方块字的汉语，却提出了很多挑战性任务有待我们解决，例如怎样划分“字”和“词”的界线（“企鹅”是两个字？一个词？还是两个词？“美利坚合众国”（The United States of America）是一个国名，在英语由5个词组成，“中华人民共和国”（The People’s Republic of China）也是一个国名，由多少个字或词组成？）这些问题每个人都可各自回答，但在语料库语言学里，则必须有一些毫不含混的规则，才能实施计算机自动化处理。和这个问题有关的是汉语怎样切分，我国语料库语言学家在现代汉语方面已经作了很多有益的探索，并建立了一些语料库，并在网上公布，以供查阅，但仅是起到一个检索工具的作用。目前好像还没有公布哪一个权威公认的频数词表，因为“字”和“词”的界线分不清，人都没有弄清楚，计算机更无能为力。一个更具挑战性的任务是汉语历史语料库，这是研究汉语历史变化的重要工具；就以书面语而言，汉语经历过春秋战国、秦汉、唐、宋、元、明、清、民国、当代（且不管甲骨文、铭文、石鼓文）等阶段，对每一个阶段都应该建立有代表性（经过抽样）的语料库，才能对汉语的变化和发展作比较。经过前人的努力，大部分典籍已经句读，但是句子（或句段）内的词却没有切分，与此有关的是汉语的词类划分，仍然是一个争议甚多（“文革”前在中国语言学界里有过一次热烈讨论）的问题。连近来出版的《现代汉语词典》、《汉语大词典》都没标出“词类”。从文献检索的角度看，我国对经典著作编制索引是有传统的，以前称为“引得”（index），燕京大学图书馆洪业（1932）就介绍过“引得”和“堪靠登”（concordance），他谈到蔡耀棠对《道德经》所编制的检索：

表1. 《道德经》中“也”字的索引行（1922）（见蔡耀堂《老解老》）

行数	索引行	行数	索引行
3	使夫知者不敢为也	55	精之至也
20	我愚人之心也哉		和之至也
24	其在道也	67	若肖久矣其细也夫
29	不可为也	76	人之生也柔弱
32	天下莫能臣也		其死也坚强
53	是谓盗夸非道也哉		万物草木之生也柔脆 其死也枯槁

由此看到，“也”作为语气助词共有10次，作为表示“并列”关系的副词有3次。洪

业还介绍过一个更大型的检索器，那就是康熙43–50年（1704–1711）由皇帝组织张玉书为首的70余人历时7年而完成的《佩文韵府》，共106卷，是1万8千页的巨著[2]。该书除对所收单字（共10,235个字分4声按韵排列）注音和解释外，还收了一些合成词和词组，并注明出典，较符合Sinclair所提出的习语原则。而这完全是手工完成的。我由此想到，像《佩文韵府》这样的经典著作还很多，如《尔雅》、《说文解字》、《方言》、《释名》、《广韵》、《辞源》、《辞海》等等。它们都可以说是经过人工预处理，我们为什么不把它们都电脑化，起码能够省掉很多检索时间，如果能够建立内部连接，对研究汉语的历史和变化就功德无限。附带的一个问题是我国的学术著作似乎有一个"不良的"传统，就是书后没有索引，西方则不然。洪业曾经指出，当年James Legge把中国古籍（其中包括《左传》）翻译到英语，Fraser & Lockhart（两人都是爵士）专门编制*An Index to Tso Chuan*，英国牛津大学出版社为之发行，Legge所译的《诗经》也有索引。但是迄今为止，中国出版业并没有以此为规范，殊觉可惜。其实只要使用Microsoft Word来编辑索引，也不很难，编者和作者都可以做，要害是页码必须对应。索引很重要，绝非多余，中国著名语言学家周法高就曾经组织一个团队来编制以王念孙《广雅疏证》为基础的《广雅索引》(1977)。周著全部都是手写影印的，因为《广雅》很多古体字、异体字，而计算机的汉语文字处理系统的造字功能当时还没有，现在用起来也很麻烦。这可能也是历史汉语语料库的一个潜在困难。

除了母语，各种外语（特别是通用性最强的，如英、俄、法、西语）也应该受到语料库研究者的关注。其中英语（美国、英国、澳大利亚、加拿大）又应该占有独特的地位，因为它不仅通用性最强，又是语料库语言学的主要发源地。这里首先应该确立的一点是英语并非中国人的母语，也没有一个包括英语的双语社区。不管先天也好，后天也好，中国人并不具备使用英语的语言能力（天性、机能），所以对英语使用中的正误、语用域、型式、习语、语义韵等判断存在很多个别差异。在我国建立的英语语料库应该有两种：一种是英语学习者语料库，它的特点是学习者英语有不同的发展和变化阶段，如小学、初中、高中、大学、研究生等等；另一种是英语使用者语料库，它的特点是：英语应该是接近英语母语使用者，其内容则随着社会和文化的变化而有所不同，如英语版的《新华电讯》、《中国画报》、《中国文学》和很多中国经典著作的英译本。以前一种而言，一个主要的问题是语料的来源，中国英语学习者只有在课堂内才接触英语，课堂外也可以接触一点，如看英语原版电影或电视剧，那也只限于接受性语言，是输入。产出性语言（书面和口语）很难获取，更不用说从大量语料中抽样。所以根据这些语料库来概括学习者的英语特点是有局限的，应十分小心。另外学习者语料库必然有很多语言使用中的失误，从发音、拼写、语法、词汇到语用都有，而这些误差频数往往是判别英语水平高低的标准。准确地说，这些失误其实包括mistakes（失检）和errors（错误），两者既有联系，也应有所区别：前者是语言运用（performance）失误，如不小心，经指出后学习者可自行改正；后者是语言能力（competence）失误，经指出后也无从改正，因为学习者还不懂（见桂诗春2005）。对学习者语料库我们虽可进行自动化词类标记（如使用Claws软件），但是因为存在失误，大大影响其标记准确性。由Granger发起的国际英语学习者语料库（ICLE，International Corpus of Learner English）就由多个国家合作收集语料组成，并没有做任何失误标记[3]。桂诗春和杨惠中（2003）所建立的中国学习者英语语料库（CLEC，Chinese Learners English

Corpus）是公开发表的带有语言失误标记的一个100万词的语料库，已为我国语料库研究者提供了方便易用的资源；但是使用者往往认为使用了这些数据就能理所当然地说明问题，而对它的研制和开发，以及所提供数据存在的问题缺乏足够了解。例如：1）CLEC收集的是书面语，但来源却很不相同，因为汉语社区缺乏使用英语的语言环境，所以写的东西并非自发性的（spontaneous）语言使用，有不少是考试中的命题作文，甚至是复述练习，即使是日记、书信也都是布置的作业。CLEC只有100万词，但因为定位在对语料作失误标记，要耗费很多人力，所以难以扩大；2）因为语言来源很不一样，原来设计的题录，有些无法填上，如性别、年龄、在读学校类型、写作时有无词典帮助等项；3）失误的标记由10几个人在不同地区完成，很难统一。更重要的是有些失误可以从不同角度来标，如冠词和名词的单复数、用语和句法等等。试看下面的一句话：Chinese young people are facing increasingly serious problem [np6, s-] on job-seeking, because of big population and less [np8, 1-] post [np6, s-]. 标记员认为有3个失误，两个是[np6]（名词的“数”），一个是[np8]（“数量”）。但是光改了这几点，句子就通顺了吗？其实这牵涉到冠词的应用，一种说法可能是problems, 另一种说法可能是the (或an) increasing serious problem, 至于后一个post则不是改为复数可以解决的，应该是few job opportunities。不管是单数还是复数，problem后面跟着的介词应该是of，而不是on。而且big population前面也要有特指，China’s 或her。又如下面的一个句子：Because of this case, people is [vp3 1-] easier to find jobs. [vp3 1-]表示动词出现一致性错误，但是改成 *people are easier to find jobs也不解决问题，应该说it is easier for people to find jobs，才较为通顺一点。

8. 您如何评价您个人对中国语料库研究发展的贡献？

我对中国语料库研究发展说不上有什么贡献，只能说在结合中国实际方面作了一些探索，我和杨惠中教授所领衔建立的CLEC，是属于早期的研究，建成后我们公开宣布这个语料库属于公共资源，可以随意采用，由此引发了一批对中国英语学习者的英语考察，最早的是我们自己的研究，见杨惠中、桂诗春、杨达复（2005），后来被采用的研究应该在百篇以上。美国、日本、新加坡、中国香港等国家和地区的学者都来了解。如上所言，CLEC也有不少有待改善的地方。

我还出过一本关于语言学语体研究的著作（2009），这是在Biber的启发下完成的，把语言学语体（ECOL，English Corpus of Linguistics）和通用型语料库（如FLOB）和BNC的科技语料（包括自然科学、应用科学、社会科学）用多特征/多维度方法来加以比较，也获得一些有用的资料和数据：从语法来说，名词化、名词、现在时、被动式、过去分词省略*wh*-式、介词、连接式、修饰方式、分裂辅助词、无人称、形态词都是把语言学语体和通用性语体区分开来的一些特征。关键性分析的结果则表明，语言学语体拥有其自身的一批专业性词汇，引导出一些搭配词群，同时对它及其他次专业词汇赋予语言学的内涵。这些词汇在定义性、分类性、分析性（包括结构性、功能性、比较性、说明性）、修饰性语言、词汇包等方面均有其语言学语体的特点。语言学语体的功能是概念性的、语篇性的、以传递和讨论信息及内容为主，它还具有抽象性（名词化、名词）、被动式、逻辑性（连

接式）、客观性（there、可能情态词、人称代词较少）、修饰性（定语性形容词、表语性形容词、普通副词、其他副词、分裂辅助词）、紧凑性（过去分词、过去分词省略*wh*-式）的特点。做这项研究的目的是建立另一个我国研究生（硕士和博士）语言学论文语料库，以作比较，从而研究他们论文写作的特点和问题。这个语料库收集了50多万词，首先是发觉它的代表性有问题，一下子难以解决，ECOL是从10个分支学科（应用语言学、认知语言学、自然语言处理、心理语言学、语用学、语义学、社会语言学、文体学和理论语言学）抽样组成的，而我国研究生的论文研究题目则集中在应用语言学和语用学两个方面；因为代表性不一样，容易产生偏颇。其次是论文写作不规范，有不少地方从原文抄录而又不加说明，所以收集的语料刻意回避"文献综述"，而集中在"讨论"和"结论"上面。我对这两个语料库的46个语法词汇特征，也曾用同样方法作过一些统计和比较，我国研究生语言学语料库有36个（78%）特征，是有显著意义差别的，其中19个（约52%）是超用的，其他是少用的。例如超用的有分类性词汇（Class, 27.92：3.341，log近似值 = 24091）、名词化（Nomil 52：37，log近似值 = 3881）等，少用的有增强语（amplifier 1.12：1.52，log 近似值 = 1541）、减弱语（downtoner 0.45：2.12，log近似值 = 1054）、模糊限制语（hedge 3.3：5.07，log近似值 = 24.9）等。这很有可能和样本来自"讨论"和"结论"部分有关：因为下结论需要条分缕析，而且避免含混。所以我未公开这些结果，以免造成误解。

9. 在您看来，从事语料库研究应具备哪些方面的学科素质？您对从事语言库研究的年轻学子有什么样的忠告？

在上面谈到学科建设的几个方面，我想也可以用来指学科素质，总之"学无止境"、"学然后知不足"，我们不应把语料库语言学看成是一门孤立的学科。它是一支箭，它本身需要磨勘，但更需要射御有术，命中目标。在射御时，既要看准目标，也要环视其周围环境，了然于心。做学问必须开拓视野，诺贝尔奖金获得者、著名认知科学家Simon曾经以有机体觅食为例，说明它的存活和视野有密切关系，如果按照他所提出的著名Q（不能存活的机会）公式计算，如果视野（v）很窄，只有1.5，而其他变量（"食物的丰富程度（p）"、"环境中的路径（d）"、"储存容量（H）"）不变，则Q = 0.897，如果v大一倍，为3，则Q = 0.286，如果再增加为4，则Q = .002。见桂诗春（2013）[4]。这就牵涉到一个不可回避的问题：要在当今时代增加存活机会，要看准目标和环视周围环境必须首先自我"定位"——我们站在什么地方？我们应该定位在"大数据时代"。

因此，我愿意向从事语料库研究的年轻学子推荐一本书，就是Mayer-Schonberger和Cukier所著的《大数据时代》（Mayer-Schonberger & Cukier 2013）[5]。书中举了几个例子说明大数据时代的到来（其中一例是2009出现甲型H1N1流感新病毒，Google把5,000万条美国人最频繁检索的词条和美国疾控中心在2003年至2008年间季节性流感传播时期的数据进行了比较：为了测试这些检索词条，总共处理了4.5亿个不同的数学模型，他们的软件发现了45条检索词条的组合，将它们用于一个特定的数学模型后，他们的预测结果与官方数据的相关性高达97%，而且判断非常及时，不会像疾控中心一样要在流感爆发一两周之后才可以做到。书中提出在大数据时代来临时需要我们改变思维方式的三个问题，我们可以结合语料库语言学来进一步思考：

1）更多：不是随机样本，而是全体数据。在大数据时代，我们可以分析更多的数据，有时候甚至可以处理和某个特别现象相关的所有数据，而不再依赖于随机采样。所以“样本 = 总体”，数据是越多越好。语料库语言学是敏锐地感到网络兴起对其影响的学科之一，因为像BNC那样现有的语料库难以适应考察英语语法的短暂发生点，而且只集中在英语世界的内环区，又覆盖不了一些新文本如博客、聊天室、交互式网上杂志等，而且网络语言可能是影响语言变化的主要信息源。进入21世纪以来，语料库语言学研究者就开始注目于怎样利用网络来推进研究；一般来说，有两大倾向：一是WaC（Web as Corpus, 把网络作为语料库）；一是WfC（Web for Corpus，用网络来建语料库）。前者是利用现成的商用搜索引擎（如Google）来进行检索，或在此基础上进行一些改进（预处理或后处理），如Google（https://books.google.com/ngrams）, WebCorp（http://www.webcorp.org.uk/live）或WaCky（http://wacky.sslmit.unibo.it）等等。后者是把网络作为信息源，从网址直接下载网页，然后借助计算机程序来建立庞大离线监控语料库。Hoffmann（2007）就介绍了怎样从CNN网页下载文本（http://transcripts.cnn.com/TRANSCRIPTS/）来建立语料库。这些探索都见于Hundt *et al.*（2007）。但是不管哪一种做法，都碰到很多尚待解决的问题，因此受到老一代语料库语言学家的质疑，如Leech（2007）。其中一个核心的问题是网络资源难以满足语料库的基本要求，所以Leech称之为“‘代表性’的圣杯”[6]。首先是网络上的资源并没有口语体，都是书面语，这难以说就是语料的“总体”，它仍然是一些有限的话语，整个网络的语料有多少也无从提供，所以有些网络语料库只是起到一个检索器的作用，无法提供一个频率的词表。而且这些语料是何人（本族语还是非本族语使用者？年龄？性别？受教育情况如何？）使用的，也不知道。语篇的长度和读者信息也无从得悉（是娱乐性的小报还是严肃的大报？），而且有些商业性搜索引擎和算法并没有公开，其搜索结果并不稳定，更不用说有很多重复资料。一般的检索也没有词类标记，这对我们了解检索词的使用也打了折扣。所以这些问题对语料库的代表性、平衡和可比性都很有影响，最后必然导致语料的偏态。在一些语料库语言学研究者的努力下，这些问题正在一一解决，但是网上的种种搜索工具当初都不是为语言学检索而设计的（特别是从召回率和准确率的角度来搜索语言特征，例如要找出由-itis组成的名词就不容易），所以目前还做不到用网络语料来代替语料库；但它可以对语料库提供更多参照性数据，有利于我们进一步观察。

2）更杂：数据量的大幅增加会造成结果的不准确；与此同时，一些错误的数据也会混进数据库。然而，重点是我们能够努力避免这些问题的出现。我们从不认为这些问题无法避免，基至需要学会接受它们。这就是由“小数据”到“大数据”的重要转变之一。在语料库越来越大的今天，这对我们研究语料也不无启发，允许不精确数据的出现已经成为一个新的亮点，而非缺点。因为放松了容错的标准，人们掌握的数据也就多起来，可以利用这些数据来做更多的事情，做多角度的探索，这不也是Biber所强调的多特征/多维度分析吗？所以我们不必拘泥于具体的频数，而需更多地注意倾向和发展方向。

3）更好：不是因果关系，而是相关关系。知道“是什么”就够了，没必要知道“为什么”。在大数据时代，我们不必非得知道现象背后的原因，而是要让数据自己“发声”。其实语料库研究把重点放在搭配（collocates）、型式（patterns）也正是在寻找相关关系，而不在于说明其因果关系。

当然，我觉得大数据时代要求使用全体数据，那就无所谓概率和随机抽样，但语料库语言学的一套运作方法都是以概率论为基础的，故有所谓probable grammar（Halliday），probable language（Newmeyer），probabilistic linguistics（Bod *et al.* 2003）这样的说法。那又怎样理解和调协这两种提法呢？我觉得Mayer-Schonberger提出的是一种目标，所以有“更多”（more）之说，而语料库语言学则是从语言现实和语言使用出发，Bod在书的《序言》里指出，“概率无所不在（everywhere）……概率渗透了整个语言系统”，类符（types）和形符（tokens）的概率都起了重要作用，一个说话人所碰到的包括特定词缀的不同词语（类符）的数量和那些词语（形符）的频数都是同样重要的。而且全球每时每刻都有几十亿人在不同的角落里使用语言，要使用其“总体”，既有困难，又无必要。所以Mayer-Schonberger & Cukier（2013）也指出，在小数据时代的随机采样是用最少的数据获得最多的信息，也是“非常有见地的”。他还说，“有些时候，我们还是可以使用样本分析法，毕竟我们仍然活在一个资源有限的时代。但是更多时候，利用手中掌握的所有数据成为了最好也是可行的选择”。所以语料库语言学在大数据时代里应该一方面保留其离线语料库，加强其代表性（而不是像Leech 所说的“只在口头上”做到代表性），另一方面是改进搜索引擎，建立以网络为基础的语料库，使它们互相补充。

注释

1. 其实按照牛津英语大辞典，把corpus当作“语料”是W. S. Allen（1956）首创，有corpus of material的说法。而Chomsky在1957年的*Syntactic Structures*也经常使用corpus这个词来说明语料和语法的关系，如corpus of sentences, corpus of utterances。他在注释里则说明*The Structure of Appearance*（Goodman 1951：5-6）就出现这样的句子：Notice that to meet the aims of grammar, given a linguistic theory, it is sufficient to have a partial knowledge of the sentences (i.e., a corpus) of the language...。

2. 因为工程浩大，参与者过百，很多注解和来源多是辗转传抄，不少讹误。

3. Granger等（Dagneaux *et al*. 1998）也曾试图对15万字的法国英语学习者（中级和高级）的语料进行失误标记，而且编制失误编辑器。

4. Simon的公式为：

$$Q=(1-p)^{(H-v)d^{v}}$$

5. 该书已译成中文，书名《大数据时代：生活、工作与思维的大变革》，译者周涛，浙江人民出版社出版。

6. Holy Grail原为（耶稣离世前使用的）圣杯，转义为“难以实现（无法实现）的梦想”。

参考文献

Adolphs, S. 2006. *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies* [M]. London: Routledge.

Baker, P. 2006. *Using Corpora in Discourse Analysis* [M]. London: Continuum.

Biber, D. 1988. *Variation across Speech and Writing* [M]. Cambridge: CUP.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Spoken and Written English* [M]. London: Longman.

Bod, R., J. Hay & S. Jannedy. 2003. *Probabilistic Linguistics* [M]. Cambridge, MA.: The MIT Press.

Carter, R. 1997. *Investigating English Discourse: Language, Literacy and Literature* [M]. London: Routledge.

Chomsky. N. 1957. *Syntactic Structures* [M]. The Hague: Mouton.

Conrad, S. & D. Biber. 2009. *Real Grammar: A Corpus-Based Approach to English* [M]: London: Pearson.

Dagneaux, E., S. Denness & S. Granger. 1998. Computer-aided error analysis [J]. *System* 26(2): 163-174.

Dryer, M. 2007. Review of Frederick J. Newmeyer, *Possible and Probable Languages: A Generative Perspective on Linguistic Typology* [J]. *Journal of Linguistics* 43: 244-252.

Fries, C. 1952. *The Structure of English* [M]. New York: Harcourt Brace & Co.

Garside, R., G. Leech & G. Sampson. 1987. *The Computational Analysis of English* [M]. London: Longman.

Goodman, N. 1951. *The Structure of Appearance* [M]. Cambridge, MA.: Harvard University Press.

Greenbaum, S. & R. Quirk. 1970. *Elicitation Experiments in English Linguistics Studies in Use and Attitude* [M]. London: Longman.

Gries, S. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction* [M]. New York: Routledge.

Gries, S. 2013. *Statistics for Linguistics with R (2nd Edition)* [M]. Berlin: Mouton De Gruyter.

Halliday, M. 1991. Corpus studies and probabilistic grammar [A]. In K. Aijmer & B. Altenberg (eds.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik* [C]. London: Longman.

Herdan, G. 1960. *Type-Token Mathematics* [M]. The Hague: Mouton & Co.

Herdan, G. 1964. *Quantitative Linguistics* [M]. London: Butterworths.

Herdan, G. 1966. *The Advanced Theory of Language as Choice and Chance* [M]. Berlin: Springer-Verlag.

Hoffmann, S. 2007. From web page to mega-corpus: The CNN transcripts [A]. In M. Hundt, N. Hasselmo & W. Bewley (eds.). *Corpus Linguistics and the Web* [C]. Amsterdam: Rodopi.

Hundt, M., N. Nesselhauf & C. Biewer (eds.). 2007. *Corpus Linguistics and the Web* [C]. Amsterdam: Rodopi.

Keller, R. 1944. *On Language Change: The Invisible Hand* [M]. New York: Routledge.

Leech, G. 1974. *Semantics* [M]. Middlesex: Penguin Books.

Leech, G. 1997. Teaching and language corpora: A convergence [A]. In A. Wichmann (ed.). *Teaching and Language Corpora* [C]. London: Longman.

Leech, G. 2007. New resources, or just better old ones? [A]. In M. Hundt, N. Nesselhauf & C.

Biewer (eds.). *Corpus Linguistics and the Web* [C]. Amsterdam: Rodopi.

Manning, C. & H. Schutze. 1999. *Statistical Natural Language Processing* [M]. Cambridge, MA.: The MIT Press.

Mayer-Schonberger & K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think* [M]. New York: Houghton Mifflin Harcourt.

McEnery, T. & A. Wilson. 2001. *Corpus Linguistics: An Introduction (2nd Edition)* [M]. Edinburgh: Edinburgh University Press.

Newmeyer, F. 2005. *Possible and Probable Languages: A Generative Perspective of Linguistic Typology* [M]. Oxford: OUP.

Oakes, M. 1998. *Statistics for Corpus Linguistics* [M]. Edinburgh: Edinburgh University Press.

Quirk, R. 1968. *The Use of English* [M]. London: Longman.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1972. *A Grammar of Contemporary English* [M]. London: Longman.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A Comprehensive Grammar of the English Language* [M]. London: Longman.

Renouf, A. 2007. Corpus development 25 years on: From super-corpus to cyber-corpus [A]. In R. Facchinetti (ed.). *Corpus Linguistics 25 Years On* [C]. Amsterdam: Rodopi. 27-49.

Sampson, G. 1997. *Educating Eve* [M]. London: Cassell.

Sampson, G. 2001. *Empirical Linguistics* [M]. London: Continuum.

Sampson, G. & D. McCarthy. 2004. *Corpus Linguistics: Readings in a Widening Discipline* [M]. London: Continuum.

Simpson, R. & J. Swales. 2001. North American perspectives on corpus linguistics at the millennium [A]. In R. Simpson & J. Swales (eds.). *Corpus Linguistics in North America: Selections from the 1999 Symposium* [C]. Ann Arbor: The University of Michigan Press. 1-14.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.

Sinclair, J. 2003. *Reading Concordances* [M]. London: Longman.

Sinclair, J. 2004. *Trust the Text* [M]. London: Routledge.

Sinclair, J. & A. Renouf. 1988. A lexical syllabus in language learning [A]. In R. Carter & M. McCarthy (eds.). *Vocabulary and Language Teaching* [C]. London: Longman. 140-158

Stubbs, M. 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture* [M]. London: Blackwell.

Svartvik, J. 2007. Corpus linguistics 25+ years on [A]. In R. Fachinetti (ed.). *Corpus Linguistics 25 Years On* [C]. Amsterdam: Rodopi. 11-25.

Thorndike, E. 1921. *The Teacher's Word Book* [M]. New York: Columbia University.

Willis, D. 2009. *The Lexical Syllabus* [M]. London: Collins ELT.

蔡耀堂，1922，《老解老·道德经串珠》[M]。作者自刊。

桂诗春，2005，中国学习者英语言语失误分析 [A]，载杨惠中、桂诗春、杨达复（编），《基于CLEC语料库的中国学习者的英语分析》[C]。上海：上海外语教育出版社。

桂诗春，2009，《基于语料库的英语语言学语体分析》[M]。北京：外语教学与研究出版社。

桂诗春，2013，向前看，向横看——略谈跨学科的必要性 [J]，《中国外语》（3）：4-8。

桂诗春、杨惠中，2003，《中国学习者英语语料库》[M]。上海：上海外语教育出版社。

洪　业，1932，《引得说》[M]。北京：燕京大学引得编纂处。

杨惠中、桂诗春、杨达复，2005，《基于CLEC语料库的中国学习者英语分析》[M]。上海：上海外语教育出版社。

周法高，1977，《广雅索引》[M]。香港：香港中国语言学研究中心。

祝启波，1991，《石油英语频率词典》[M]。北京：石油大学出版社。

通信地址：510420 广东省广州市广东外语外贸大学外国语言学及应用语言学研究中心

英语搭配框架的意义单位再探

华南师范大学 何安平

提要：本研究探讨语料库语言学短语视角下搭配框架的意义单位模式，尤其关注意义单位的起止边界以及框架两端语法词对其短语序列的意义制约。基于Brown和LOB书面语料库中the ? of、a ? of和the ? in三个搭配框架约2.5万个实例，本文采用“延伸式搭配框架”的视角和“累积频数”的驱动方式构建了三者各自的词汇语法共选模式。结果发现这类“冠词+名词+介词+名词”的短语序列：1）在语法结构上有层层镶嵌、循环拓展的倾向；2）在语义上对框架目标名词的描述有不同语义偏好，显示框架两端的语法词对其意义模式颇有制约；3）在功能上起命题标识和语篇衔接作用，前两个搭配框架尤甚。

关键词：搭配框架、语料库、延伸式框架、累积频数、意义模式

1. 引言

搭配框架（collocational framework）较早由Renouf & Sinclair（1991：128）定义为“由两个非毗邻的“语法词”共同搭建一个框架而让其他词汇有选择地填入”，例如a ?[1] of和in ? with。后来有研究者（如Eeg-Olofsson & Altenberg 1994；Marco 2000）将搭配框架分为“语法词 ?语法词”（如as ? as）、“实义词 ?语法词”（如many ? of）、“实义词 ？ 实义词”（如wait ? minute）等多种搭配序列，而且搭配框架中间的填充词也从一个扩展至两个或多个（如 a ? ? of → a large number of）。此外，还有Philip（2008：97-99）把搭配框架归属于短语构架（phraseological skeleton）范畴，内含搭配框架、词汇语法框架（lexicogrammatical frame）和半预制包词组（semi-prepackaged phrase）三种。Stefanowitsch & Grices（2008: 933）把搭配框架归属于结构敏感搭配（structure-sensitive collocate）范畴，内含搭配结构（collostruction）、语法型式（grammar pattern）等。Gries（2008：20）把搭配框架和类联接视为三大类短语之一，将其置于连续的n元短语（n的上限为5，如a very large number of）和连续的词性标记（如Adj N、N N、N P N等）之间，并且提出了搭配结构分析（collostruction analysis）的概念。本文聚焦Renouf & Sinclair（1991）定义下的搭配框架，针对的是英语中高频出现的语法词（或称虚词，例如冠词、介词、连词等）。这些语法词的使用频数通常会占语篇总词量的一半以上，而且往往成对出现，构成了英语“最普遍的共选现象”（Sinclair 1999：157）。然而“它们的本质、功能以及在语言理论中的地位一直未得到充分认识。所以很有必要研究这种现象的规律或模式”（Renouf & Sinclair 1991：129）。概而述之，本文关注的是搭配框架的意义单位构建。

2. 理论依据

关注搭配框架的意义单位主要受两位学者的启发。一位是20世纪中期伦敦学派的领

军人弗斯，他提出了语境模式论（contextual patterning）（Firth 1957：35），他认为“所有对意义的陈述其实都是陈述词语在语境中的相互关系（contextual relations）”（Firth 1957：11），即“通过语言学技术或术语对反复出现的词语特征进行模式化的系统陈述”（Firth，转引自Palmer 1968：187），而“作为语境一个组成部分的语篇以及词语在该语篇中形成的语境模式都共同对该词语的意义起决定性作用”（Firth，转引自Tognini-Bonelli 2001：4）。另一位是继承和发展弗斯语境论思想的语料库语言学领军人Sinclair（Sinclair 1991，2004：31-39）。他提出意义单位论，认为语言中的意义单位主要是短语而非孤立的单词。短语是构建意义模式的基础，意义模式的内涵包括一个核心项和四层共选关系，即搭配、类联接、语义偏好和语义韵。其中核心项可以是单个词、多字语或是框架；搭配指核心项与周边词在词汇层面上的共选；类联接是核心项与典型搭配词在语法层面上的共选，语义偏好指核心项与典型搭配词，或搭配词之间在意义层面上的共选；语义韵属于传递话语态度和隐性意义的语用功能（Sinclair 2002）。支撑这一意义构建模式的数据来自语料库大批量检索得到的语境共现行，它让人们前所未有地探究语言中大量反复出现的词汇共选现象，该现象深刻地揭示“语言使用的短语倾向性”本质（Sinclair 2004：29-31）；但是目前该项目研究也存在“难以廓清短语单位起止边界的‘短板’”（何安平 2013：70）；加之以往对意义单位研究的对象大都是连续性的词语组合（如拓展词项、机切语块），所以正如Sinclair所（2004：39）指出的，我们“有必要将短语单位的研究延伸至单词之外的各种变体。其中很值得进一步研究的就是那些由语法词而非实义词为核心的搭配框架”，“有必要研究它们的模式”，“尤其要廓清制约搭配框架的因素”（Renouf & Sinclair 1991：129，133）。

3. 前人研究综述

互联网搜索显示，专门研究英语搭配框架意义模式的论文并不多见，但以下几项成果却对本研究很有启发。首先是Renouf & Sinclair（1991：136）早年对搭配框架an ? of中间的填充词与框架后续词的语义关系进行的开拓性研究。他们提取了1.1千万词次的伯明翰英语语料库中an ? of的of之前和之后的、频数为10次以上的名词（简称N1和N2），然后对它们进行语义功能分类。结果发现了两种语义关系：一是N1与N2是从属语义关系，包括N1是对N2的量度（如an ounce of gold）、聚焦（如an edge of the table）、支撑（如an act of murder）；二是N1与N2是构建命题的并列关系（如an extension of the course →the course extends）。而且“这种单位还可能扩充，成为一个更大的短语单位的一部分，如Verb + Obj + with an air of + Noun”（同上）。

Diniz（2007）的研究进一步论证搭配框架的意义单位属性。她从共约130万词次的英语教材话语和教师课堂话语中，提取了1.25万例the ? of，并对中间填充词的语义和功能进行分类，发现最高频的类别均是表达“本质类”的抽象名词（如nature、basis、concept），其功能是凸现这两类学术话语中具体命题（即of的后续名词）的本质性、核心性和重要性。她还从语音层面上发现the ? of在口语里基本是一个不带停顿的语音单位，从多个层面证实“词语总是倾向以共选的方式组成固定或半固定的短语形式，传递着语言使用者的多种交际功能”（Diniz 2007：144）。

Marco（2000）的搭配框架研究关注其语篇意义和语域特色。她分析了近30万词次的医学英语论文语料中3种最高频的搭配框架的语篇功能和语义语法序列。其中the ? of主要用于构建表述度量和医疗过程的名词化短语（如the degree of，the addition of）；a ? of用于量化和分类性描述（如a proportion of、a history of）；be ? to用于构建因果类和相似性的关系（如be due to、be likely to）。研究发现某些普通词汇，如start进入the start of框架就成了医学专业的短语单位，因为它百分之百地作为名词用于表述某种治疗或试验的开始。另有一些半专业性词汇（如cause）则主要通过the cause of框架用于表述病因或疗效。还有一些搭配框架与其前后的短语构成更大范围的序列结构，来表达该专业的话语模式，并且与该类语篇的具体话步或篇章功能直接挂钩。例如the result of往往与前面的based on或derived from等“来源”类动词短语形成共选序列，用于表述当前话题与前人研究的相关性和延续性，而且大都出现在该类语篇的“文献综述”部分等等。

Warren（2011）认为搭配框架是英语短语的三大形态之一（其余两类分别是意义迁移单位和组织框架，详见何安平2013）。他一方面运用Hunston（2002：154）的局部语法（local grammar）理念，构建了any...may和may...any两个搭配框架的不同语法和语义模式；另一方面又参照Halliday（2009）关于当代语言演变现象的论断，即语言技术化的演变导致语言表达越来越抽象，语法的隐喻化导致越来越多名词化短语的演变，对比了科技论文、政治演说和日常会话三种语料（各5千词次），结果发现the ? of的使用频数分别为105、96和11，从而揭示搭配框架是比较正式的书面语语体特征，而且与语言的名词化发展倾向相辅相成。

以上对于搭配框架的意义研究从短语拓展至语篇和语类文体，但却少有专门在语料库语言学的意义单位视角下把搭配框架的前面、中间和后面的搭配词组合起来作为一个完整短语序列来探究其多层意义模式；而且“对于这类多词单位的意义模式的起止边界也未有统一界定”（Danielsson 2007）。基于前人对搭配框架的研究成果及其尚存的研究空白，本研究尝试探讨以下问题：

（1）如何基于搭配框架构建多层意义模式？

（2）搭配框架意义模式的边界有何特征？

（3）搭配框架两端的语法词对其意义模式有何制约？

4.研究的思路与方法

首先基于英美两大经典语料库Brown和LOB（共200万词次）探讨the ? of作为短语单位在书面语语料库中的典型意义模式，尤其关注意义模式的边界特征问题；然后将该模式与这两个语料库中的另外两个搭配框架，即a ? of 和the ? in的意义模式进行对比，从而探讨搭配框架两端的语法词对意义模式构建的制约。

在研究方法上，本研究受语料库语言学的“局部语法”理念和短语意义单位构建模型的启发，尝试将上述三个搭配框架分别作为意义单位的核心项，并从词汇、语法、语义及语用功能等多个层面构建其意义模式。首先是对意义单位的边界设定，这里汲取Diniz（2007：73）“延伸式搭配框架”（extended framework）的理念，将the ? of视为可延伸的N

词序列，即“... + the + ？ + of + ...”。同时参照Hunston（2008：272）和Danielsson（2007）的“累积频数”（cumulative frequency）方法，即“不计较词语序列在语料库中的绝对频数，而关注序列中每个局部构成成分的频数特征”[2]，从框架的中间词开始先向右然后向左依次提取框架内（即MC[3]）和框架外（R1、R2... L1、L2...）的共选词项中最高频[4]搭配词的词性，以构建意义单位的语法结构序列（类似语料库的“类联结”）；然后对最高频词性中的词汇进行语义范畴归类，以构建意义单位的语义序列（语料库的“语义偏好”）。最后借鉴“局部语法”的理念，即“关注语言中某一类意义表达的局部形式特点和词汇语义特点”（Hunston 2002：154），基于上述序列诠释搭配框架拓展形式的意义功能。

由于仅从语法结构上就可预测上述三个搭配框架的中间词都是名词（以下缩写为the N of, a N of和the N in），所以可直接从已有词性附码的Brown-TAG和LOB-TAG语料库中提取，如设置the_at* *_n* of_in提取the N of的语境共现行；然后按R1->R2->L1->L2依次检索，以前面一个位置频次最高且等于或超过该位置搭配词总数30%的词性类别为基础，逐步向框架两端伸延，直到在最外向的位置上查不到频次超过30%的词性类别为止，以此作为探究搭配框架意义模式并奠定局部语境宽度的基础。下面以the N of的累积频数提取过程为例。

1）在MC提取到18,583例the N of，接着向R1伸延，获15,722例the N of + ?，R1的最高频词性为冠词类（简写为At，6,101例，占39%），其次为名词（5,132例，占33%）

2）向R2伸延，提取到5,024例the N of + At + ?, R2的最高频词性为名词类（简写为N，3,426例，占68%）

3）向R3伸延，提取到2,073例the N of + At + N + ?, 此时R3再没有一类词性的频次 ≥ 30%；于是转向L1伸延，提取到2,631例？ + the N of + At + N，L1的最高频词性为介词类（简写为Prep，1,600例，占61%）

4）向L2伸延，提取到1,312例？ + Prep + the N of + At + N，L2的最高频词性为名词（569例，占43%），其次为动词（392例，占28%）

5）向L3伸延，提取到435例？ + N + Prep + the N of + At + N，L3的最高频词性为冠词（138例，占32%）

6）向L4伸延，提取到124例？ + At + N + Prep + the N of + At + N，L4的最高频词性为介词（60例，占48%），形成Prep + At + N + Prep + the N of + At + N的语法序列，此时的词性序列已出现“介+冠+名”的循环排列，且数据稀薄（仅百例上下），故不继续伸延。换句话说，延伸式搭配框架已走到频数意义上的结构边界。

5. 语料库分析与结果讨论

采用上述方法检索出the N of、a N of和the N in共约2.5万例，得出三个延伸式搭配框架的首个语法结构序列如下：

(Prep + At)[5] + N + Prep + the N of + At + N（569例）

(Prep + At) + N + Prep + a N of + N（201例）

(At + N) + Prep + the N in + At + N + (Prep)（201例）

对比以上三个序列可看出每个搭配框架都各自带出了一个含介词及宾语的名词短语（即 the/a + of/in + At + N）；而该名词短语又置于一个更大的介词短语中（见三个框架之前的Prep）；该介词短语往往再与前面的另一个介词短语（即Prep+At+N）相连，从而形成一种循环结构的模式，而该结构模式的边界大致是8 至9 个词位。这一发现表明以“冠词+名词+介词”为特点的搭配框架往往有层层相嵌的倾向，这种延伸式的搭配框架证实了Renouf & Sinclair（1991：140-141）所说的“搭配框架呈连贯式的流线型延伸”，“由于高频语法词总是与实义词间隔着出现，其形成的模式往往是一个搭配框架的末端成分成了另一个搭配框架的起始成分”，“人们对语言单位的选择往往是多个词而非单个词的这种模式再次证明与以往‘空位-填充式’语法模式有很大不同的短语倾向”，“搭配框架成为语言产出的主流，而语法却处于适应它的地位”。

接下来探究这三个相近似语法结构序列在语义上有何异同。曾有学者指出在构建意义单位过程中，“当从词汇实体（如搭配）逐步过渡到类联结和语义模式时，必须逐渐淡化前后词位排序并且更加包容各种变体”，“在进行语义偏好层面的分析时，对搭配词的语义归类不必拘泥于其词性或语法位置，而要关注其语义取向”（Sinclair 2004：42，33），尤其要“关注其中的实义词类在语义上的共享性”（Stubbs 2002）。所以本研究在对上述三个语法结构序列中的实义词（全是名词）的语义偏好进行分析时，也首先对这些序列的核心部分（即不带括号的部分）的实义词（全是名词）进行语义归类。具体做法是提取最高频的10个MC名词并观察其语义偏好，然后基于这10个高频词继续使用累积频率的方法进行框架之后（即1R-2R）和框架之前（即1L-2L)的实义词语义归类，结果见表1。

表1.　三个搭配框架中频数最高的10个核心搭配词(MC)

频数排序	the N of	a N of	the N in
1	end	number	men
2	part	couple	people
3	rest	group	bush
4	top	period	atoms
5	bottom	variety	dust
6	center	sense	floor
7	head	series	workers
8	edge	state	warden
9	back	lot	successor
10	foot	piece	shelter
占MC总词数比例	102/569 = 18%	59/201 = 29%	18/122 = 16%
占MC词类总数比例	10/357 = 3%	10/132 = 7%	10/110 = 9%

如表1所示，各框架的10个最高频名词虽然在词的种类（type）分别只占总数的百分之3至9，但其频次（token）却占总数的16%至29%，故有一定的代表性。其中the N of的中间词语义范畴明显是表述方位和时间的“时空类”词；a N of则明显是表述数量或量度单位的“度量类”词；而the N in却多用于表述人和事物等日常普通词。可见三个框架的中间核心词各自有不同选择倾向。进一步探究制约这些选择倾向的因素驱使我们基于这三组高频名词去提取各自框架之后和之前的搭配实义词，结果见表2。

表2.　基于10个高频名词提取的左右分别排行头10位的搭配词

搭配词	右边搭配词			左边搭配词	
搭配框架	the N of	a N of	the N in	the N of	a N of
1	stairs	years	village	way	balance
2	screw	days	vincinity	interest	public
3	pin	mind	China	use	testimony
4	patrol	men	county	top	shares
5	family	time	land	tendency	frequency
6	face	people	hall	stage	anecdote
7	book	youths	bank	size	ministry
8	day	underwriters	body	side	frame
9	bank	terror	unit	settlements	issues
10	year	suits	face	self	association
语义归类	日常普通词占74%[6]	时间和人、物词占51%	地点类词占78%	抽象词占45%	抽象词占58%

表2显示在右边搭配栏目，the N of后面的名词多为日常生活的普通词汇；a N of的右边词主要表述是时间，人和物等，而且半数以上是复数名词；而the N in的右边词却明显是地点或单位的词。至于框架左边的搭配词，the N of和a N of各有近半数或六成是抽象名词，the N in则因数据稀薄而未在本研究之内。综合起来，三个框架的语义序列分别是：

the N of→　抽象类词+（Prep）+ the +时空类词　of + the+ 日常普通词
a　N of→　抽象类词+（Prep）+ a +量度类词　of +时间和人、物词
the N in→　（Prep）+ the +日常普通词 in + the+ 地点类词

communication to the rest of the face
consciousness on the part of the agent
cries at the end of the valley
delight on the part of the subject
deposit on the bottom of the pan
diligence on the part of the assessors
distance from the top of the block
eruption at the back of the stage
fantasy at the end of the program
adaptation on the part of the worker

evidence from a number of people
eyes with a lot of red
frame around a series of experiences
information from a number of organisations
inspection of a number of schools
values during a period of years
judgment of a number of experts
licence for a period of years
man with a couple of kids
plans in a number of places

the people in the country
the bush in a land
the men in the ranks
the atoms in the unit
the dust in the vicinity
the people in the village
the dust in the china
the men in the hall
the bush in a boxcar
the workers in the bank

图1. 三个语义序列的实例截图

以上三个序列及举例（图1）分别表明在英语书面语中，the N of、a N of和the N in后面所带出的名词短语都有一个典型功能，就是对某个目标进行特别的描述或界定。其中the N of和a N of的目标在of之后的名词里，前者侧重目标的时空方位特点，后者侧重目标的计量特点；the N in的目标在in之前的名词里，侧重的是目标的地点位置或所属。the N of和the N in后面的名词基本带定冠词the，表示对这些名词有特别的指向和选择，形成the N+介词+the N的模式；而a N of则无此倾向，这可能与框架左端的冠词（the和a）的制约有关。而the N in后边的地点类词则可能与其框架右端的介词in制约有关。由此表明搭配框架两端的语法词都对框架的意义模式有实质性的影响。

至于the N of和a N of左边搭配词的抽象词类倾向，似可从抽象词的标识功能进行诠释。Flowerdew（2003：329）曾指出有一类“标识类名词”（signalling noun），其意义只能在它后边的语境共现中被细化或清晰化，故起到衔接前后语篇的作用。例如本文截图中的eruption at the back of the stage和evidence from a number of people实质上是标识或廓清了eruption和evidence的信息。由于这些抽象词都处于搭配框架之外的更大一个框架中（见前面的语法结构序列），所以这种延伸式框架也可以视为一种“壳块”（shell bundle，详见Diniz 2007：122），其主要作用是通过反复出现的局部结构模式来构建局部的命题或概念（temporary concept-formation，详见Schmid 2000）。这样一来，仅由虚词构建的搭配框架就构成了具有意义功能的短语单位。

6. 小结

Renouf & Sinclair（1991：141）曾指出“研究搭配框架的最大意义在于挖掘它们在语言使用中的能产性潜力”。本研究尝试构建三个不同搭配框架的意义模式，首先从形式结构层面使用“累计频率”的方法穷尽框架的延伸边界，结果发现这类“冠词+名词+介词+名词”的搭配结构序列往往会嵌入层层循环状态，从一个侧面证实语言使用的块状性和延续性。接着基于语法结构序列的语义模式探究，发现搭配框架两端的语法词均对意义单位的构建有明显制约，表现为对框架的中间及左右的词有明显的语义选择倾向，而绝非是开放式的、允许符合词性的任何词填充。Jeffries（1998）曾指出制约语言意义的因素“有一些很仰仗语法，而其他的则显然要靠语义”，本研究发现就搭配框架的意义而言，这两者都缺一不可。

另外，本研究还发现搭配框架的延伸式语法语义模式会形成“壳块”，并在构建语篇局部概念命题和衔接语篇单方面发挥作用。它一方面表明现代英语书面语的名词化短语倾

向，另一方面也揭示了短语有很强的语篇衔接功能。

最后，本研究对英语教学的启示在于它揭示“词的共选并非因为它们自身的频繁出现，而是因为他们已经成为高频短语中的一个组成部分。语法词也并不单独出现，其功能在于构建一个更大的单位”（Stubbs 2002：235）。虽然目前很多教师开始重视诸如the end of the、the edge of the、the bottom of the等词块教学；但同时还有必要将诸如the * of the这类搭配框架视作一个整体，甚至看做可延伸的意义单位而进行更多层面语法和意义功能解读，从而训练学生在“壳块”层面上构思英语（尤其是书面英语）的概念和命题表达。

注释

1. ？在本文表示一个空缺的词，下同。

2. 这两位学者都曾用此方式探究核心项为单个词的意义单位，例如，4.5亿词次的BoE语料库中虽然只有43例will have to decide whether to，但它是从29, 731例decide中按L1->L2->R1->L3->R2顺序依次提取每个位置中最高频的搭配词组合而成，所以依然成为值得注意的意义单位。杨素香、何安平（2013）曾尝试用此理念探讨the ? of，但本研究将该方式应用得严谨。

3. 本文的MC（mid-collocates）指框架的中间填充词，R1和L1分别指框架右端，即of后面第一个位置和框架左端，即the前面第一个位置的搭配词，R2，L2照此类推。

4. 本文对“高频”的界定参照Stubbs（2002：216）在批量提取英语短语的核心词与周边词搭配强度要等于或大于搭配词总数30%的比率；之后对“语义序列”的最高频语法范畴也按其内含的词语频数占据跨度之内全部词频的30%及以上这一设定。

5. 序列中带括号的部分表示数据稀薄，左右末端表示之后再没有任何词性类别等于或超过该位置所有搭配词的30%。

6. 表中的百分比是这类语义词占该部位词频总数的比率。

参考文献

Danielsson, P. 2007. What constitutes a unit of analysis in language? [OL]. *Linguistik Online* 31. http://www.linguistik-online.de/index.html (accessed 20/6/2014).

Diniz, L. 2007. Highly frequent function words in the light of the idiom principle: The case of “the” [D]. Unpublished dissertation. Atlanta: Georgia State University.

Eeg-Olofsson, M. & B. Altenberg. 1994. Discontinuous recurrent work combinations in the London-Lund Corpus [A]. In U. Fries, G. Tottie & P. Schneider (eds.). *Creating and Using English Language Corpora* [C]. Atlanta: Rodopi. 63-77.

Firth, J. 1957. A synopsis of linguistic theory 1930-55 [A]. Reprinted in F. Palmer 1968. *Selected Papers of J. R. Firth 1952-1957* [C]. London: Longmans, Green and Co., Ltd. 168-205.

Flowerdew, J. 2003. Signaling nouns in discourse [J]. *English for Specific Purposes* 22: 329-346.

Gries, S. 2008. Phraseology and linguistic theory: A brief survey [A]. In S. Granger & F. Meunier

(eds.). *Phraseology: An Interdisciplinary Perspective* [C]. Amsterdam: John Benjamins. 3-26.

Halliday, M. 2009. Language in Science and the Humanities [R]. The Hong Kong Polytechnic University, Hong Kong.

Hunston, S. 2002. *Corpora in Applied Linguistics* [M]. Cambridge: CUP.

Hunston, S. 2008. Starting with the small words: Pattern, lexis and Semantic sequences [J]. *International Journal of Corpus Linguistics* 13: 271-295.

Jeffries, L. 1998. *Meaning in English: An Introduction to Language Study* [M]. Houndmills: Palgrave.

Marco, M. 2000. Collocational frameworks in medical research papers: A genre-based study [J]. *English for Specific Purposes* 19(1): 63-86.

Palmer, F. 1968. *Selected Papers of J. R. Firth 1952-59* [C]. London: Longman.

Philip, G. 2008. Reassessing the canon: “Fixed” phrases in general reference corpora. In S. Granger & F. Meunier (eds.). *Phraseology: An Interdisciplinary Perspective* [C]. Amsterdam: John Benjamins. 95-108.

Renouf, A. & J. Sinclair. 1991. Collocational frameworks in English [A]. In K. Ajimer & B. Altenberg (eds.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik* [C]. Harlow: Longman. 128-143.

Schmid, H. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition* [M]. Berlin: Walter de Gruyter.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.

Sinclair, J. 1999. A way with common words [A]. In H. Oksefjell (ed.). *Out of Corpora* [C]. Amsterdam: Rodopi. 157-180.

Sinclair, J. 2002. Language, Meaning and the Corpus [Z]. Sofia University of Bulgaria.

Sinclair, J. 2004. *Trust the Text* [M]. London: Routledge.

Stefanowitsch, A. & S. Grices. 2008. Corpus and grammar [A]. In A. Lüdeling & M. Kytö (eds.). *Corpus Linguistics: An International Handbook* [C]. Berlin: Walter de Gruyter. 933-952.

Stubbs, M. 2002 Two quantitative methods of studying phraseology in English [J]. *International Journal of Corpus Linguistics* 7(2): 215-244.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [M]. Amsterdam: Benjamins.

Warren, M. 2011. Extending our understanding of phraseology [R]. Keynote presentation at CLIC Beijing, China, 19-20 November.

何安平，2013，国外语料库语言学视角下多形态短语研究述评 [J],《当代语言学》(1): 62-72。

杨素香、何安平，2013，搭配框架、意义单位与大学生英语论文写作 [A]。载何安平（编),《语料库短语理念及其教学加工》[C]。广州：广东高等教育出版社。

通信地址：510631 广东省广州市华南师范大学外国语言文化学院

语料库、平义原则和美国法律中的诉讼证据

北京外国语大学 梁茂成

提要：语料库不仅可以服务于语言研究，还可以服务于其他邻近学科。在美国法庭庭审中，语料库用于协助“平义原则”的操作，已经成为一种趋势，而在提供诉讼证据方面，语料库也大有用武之地。本文通过实际案例介绍语料库在美国法庭庭审和诉讼证据提供方面的应用，以期对我国法律界和语言学界研究者和从业者有所启示。

关键词：语料库、美国法律、平义原则、诉讼证据

1. 引言

随着各语种大型语料库的陆续建成和语料库分析技术的不断更新，利用语料库所进行的语言研究越发深入，所涉及领域也得到了前所未有的拓展。梁茂成（2014）认为，经过多年的发展，语料库语言学所涉及的领域现主要包括以搭配研究为重点的核心圈、以基于语料库的词汇、语法、语篇等研究为主体的中间圈和以语言教学研究等为主要内容的外围圈。不仅如此，语料库及其分析技术还越来越多地用于情感分析（sentiment analysis）、数据挖掘（data mining）、信息提取（information retrieval）等语言学交叉学科。语料库在语言学研究及相关学科中的应用大致如图1所示。

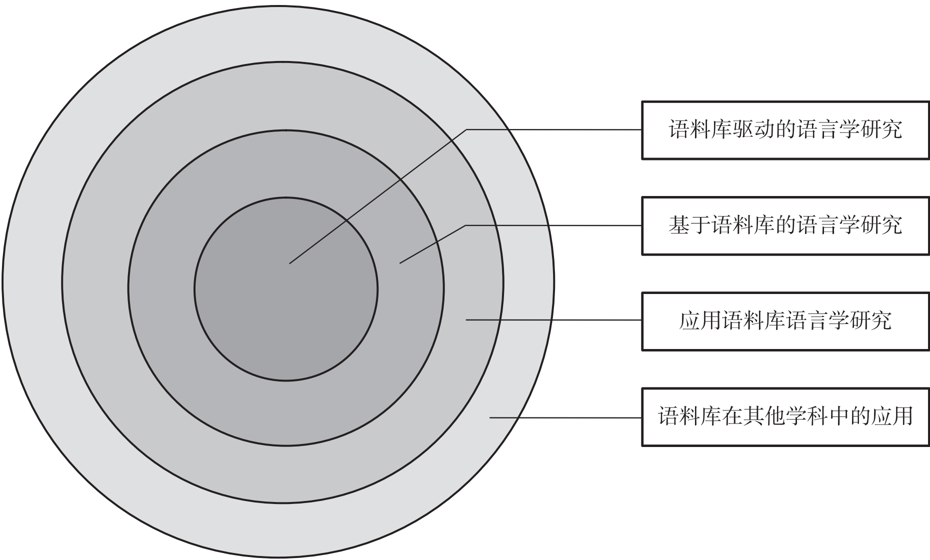


图1 语料库应用领域示意图

如图1所示，处于核心圈的语料库驱动的语言学研究以搭配研究为重点，排斥语料库语言学之外的理论预设，试图以全新的理念和方法描述语言，目前已经形成了一套便于操作的研究方法。处于核心圈外围的是基于语料库的语言学研究。在这一种研究范式中，语料库仍然被视为可靠的语言研究数据源，但该研究范式并不排斥其他理论预设，常常用于验证已有的理论假设，且并不排斥其他类型的数据源。有关以上两种研究范式的渊源及异同，参见梁茂成（2012）。由以上两层核心圈继续向外，第三圈是试图将语料库应用于语言教学的应用语料库语言学研究。这类研究属于应用语言学研究范畴，试图将语料库及语料库语言学理念与语言教学结合起来。因语言教学的需要，研究中常常可以因材施教地为处于不同学习阶段的学习者提供不同难度的语料，因而语料库的代表性不再是研究中关注的首要问题。处于最外层的是语料库在语言学外围学科中的应用，主要涉及一些交叉学科。在这些学科中，语料库只是众多可用数据源中的一种，人们注重“有用的就是好的”。纵观全图，从内圈到外圈，各种研究范式对语料库的依赖程度依次降低，对语料库语言学之外已有理论的接受程度依次递增。

本文主要涉及图1中的最外圈，主要内容包括：1）介绍美国法律体系中的“平义原则”，并通过几个典型案例介绍语料库在“平义原则”操作中的应用；2）通过具体案例介绍美国学者如何利用语料库语言学研究方法提供诉讼证据。笔者希望通过本文的介绍，引起国内语言学界和法律界学者和从业人员的重视，必要时借助语料库寻求诉讼证据。

2. 搭配研究与美国法律中的“平义原则”

2.1 搭配与意义研究

古今中外，所见略同者不乏其人。据《孔子家语》（王国轩、王秀梅 2009：136）记载，孔子曾说：“不知其人视其友”，其大意是说，物以类聚，人以群分，如果不了解某人，看他结交什么样的朋友，也就大概知道他是什么样的人了。后来，司马迁在《史记》之《张释之冯唐列传》中曾引用过孔子这句话。无独有偶，据*The Wordsworth Dictionary of Proverbs*（Apperson 1993：365）记载，英语中有一句十分类似的谚语，即A man is known by the company he keeps，讲的是相同的道理。据*The Wordsworth Dictionary of Proverbs*记载，这一说法最早出自M. Coverdale翻译的瑞士激进派改革家Henrich Bullinger的著作*Christian State of Matrimony*（1541）中，后来，Henry Smith于1591年在其作品*A Preparative to Marriage*中再次说过类似的话，到1672年这句话已经广为流传，成为英语中的一句谚语。也有文献认为，这句谚语源自拉丁语短语*Noscitur a sociis*，表达的也是“不知其人视其友”的意思。由此可见，A man is known by the company he keeps是多种文化背景下人们的共同认识。

根据Whitsitt（2005）的研究，英国作家Henry Fielding可能是第一个将“不知其人视其友”这一说法中“人”与“友”之间的关系与语言中词与词之间的关系进行类比的人。根据Glanville Williams在*Learning the Law*一书中的描述，Henry Fielding在其小说*Tom Jones*中首次根据拉丁语短语*Noscitur a sociis*创造了A word may be known by the company it keeps这一说法。

在当今的语料库语言学研究中，You shall know a word by the company it keeps（Firth 1957a：179）这句话被广泛引用，已成为语料库语言学“习语原则”的理论基础。我们无从考证Firth这句话是否与Henry Fielding的作品有关，但至少有两点是可以肯定的，即：1）Firth巧妙地借助谚语，将词语搭配与朋友结伴进行类比；2）Firth是第一个在语言学研究的语境中使用这句话的学者。在Firth看来，dark的意义之一表现在它可以与*night*进行搭配，而night的意义之一也表现在它可以与dark搭配（Firth 1957b）。换言之，意义并不存在于孤立的“词”中，因而脱离语境谈论“词义”是行不通的，通过搭配研究意义（meaning by collocation）才是正确的方法。令人遗憾的是，在Firth所处的时代，语料库建设尚处于萌芽阶段（Quirk于1959年开始建设*Survey of English*），Firth也不可能利用大型语料库来研究搭配和意义。

到了20世纪60年代，在Sinclair的领导下，伯明翰大学的语料库建设如火如荼地开展。随着大型语料库的建成，通过语料库研究搭配逐渐成为可能，这也使得Firth有关搭配的思想有了可靠的数据支撑。Sinclair吸收并进一步发展了Firth的理论，提出了“习语原则”，将搭配在语言研究中的地位提到了前所未有的高度，主张“相信文本”（trust the text），呼吁语言研究者要重视the phrase, the whole phrase, nothing but the phrase（Sinclair 2008）。至此，以搭配为研究重点成为语料库语言学最重要的特点之一。在Firth的理论体系中，意义处于语言学研究的中心地位，而词汇意义（lexical meaning）研究可以在语音、词汇、形态、句法和语义五个维度上展开。Sinclair沿袭了并进一步发展了这一思想。在他的理论体系中，语言学研究的目的仍然是意义的研究，但在Sinclair看来，意义研究的主要方法在于对搭配的有效分析。

十分巧合的是，Firth和Sinclair可能都未曾想到，若干年之后，随着一些大型语料库的建成，美国人竟不知不觉地将他们有关搭配研究的思想应用于法庭之上，通过搭配确定词义成为当今美国人执行“平义原则”过程中的重要手段。

2.2　美国法律中的“平义原则”与“平义”的确定

不言而喻，对法律条款真正意义的解读十分重要，可以直接影响到法庭的裁决。美国对法律法规的解释素以“平义原则”为基础。所谓“平义”（plain meaning），即语言共同体内普遍接受的意义，亦即普通人士在日常生活中使用词语时所指的含义。“平义原则”要求以“平义”来解释法律法规中的词语，将法学与语言学紧密地联系了起来。“平义原则”的广泛采用使得很多法律问题最终落实到对法律条款中关键词语的解读上，使得法律问题变成语言学问题，甚至变成具体的语义问题。

长期以来，美国法庭上对法律条款的解读产生分歧时，常常以词典为依据，从权威词典中查询法律条款中有关词语的“平义”，并依此裁定被告人是否有罪。由于几乎所有词典都逐项给出词条的多个可能释义，且这些释义常常脱离语境，庭审中依赖词典的做法十分难以操作，因而受到了一些法律专家和语言学家的批评。随着语料库语言学的发展，人们越发感到词典所提供的释义过于笼统，很难满足法庭的需要，开始从语料库中寻求帮助。

最早提倡借助语料库来解决法律条款中“平义”问题的学者是佐治亚州立大学法学

教授Clark Cunningham和著名语言学家Charles Fillmore。他们1995年合作发表了Using common sense: A linguistic perspective on judicial interpretations of "use a firearm"。在这篇论文中，他们从"美国诉斯密斯案"（United States v. Smith, 508 U.S. 223, 1993）等案例入手，指出语言学在法庭审判中可能的用途。在该案中，斯密斯涉嫌以枪支交换毒品。根据《美国法典》（United States Code）第18篇第一部分第924章，"任何人，在暴力犯罪或贩毒过程中……，若使用或携带枪支……，除按照暴力犯罪或贩毒罪处以相应惩罚外……，还将处以5年以上监禁"（any person who, during and in relation to any crime of violence or drug trafficking crime..., uses or carries a firearm...shall, in addition to the punishment provided for such crime of violence or drug trafficking crime, be sentenced to a term of imprisonment of not less than 5 years）。该案在庭审过程中，焦点问题在于对use（使用）一词的理解上。斯密斯认为，《美国法典》中的use指的是将枪支"作为武器来使用"，而法庭则认为，这里对use无任何界定，按照词典释义，指的应该是任何方式的"使用"。由于斯密斯将武器作为交换物品使用，法庭裁定，斯密斯在贩毒罪行的基础上获得30年监禁的额外加刑。为了证明法庭的裁决缺乏依据，Cunningham和Fillmore在研究中从英国国家语料库、美国报纸语料库等大型语料库和《美国法典》内部寻找众多例证，证明以上条款中的use指的是"将枪支作为武器来使用"，因此庭审中法官对法律文本的理解是错误的。此外，Cunningham和Fillmore在研究中强烈呼吁法律从业人员不要过多地依赖词典，而应该从语言使用实例中寻找证据。

2011年2月，杨百翰大学（Brigham Young University）学者Stephen Mouritsen发表了The dictionary is not a fortress: Definitional fallacies and a corpus-based approach to plain meaning一文。Mouritsen师从著名语料库语言学研究者，熟知语料库语言学研究方法。他在文中指出，从20世纪60年代到21世纪初，美国最高法院裁决记录中ambiguous、ambiguity（模棱两可）、plain meaning（平义）、ordinary meaning（常用义）等词语的使用频率增长了几倍甚至十几倍，充分说明法律条款的准确解读已经成为庭审中十分棘手的问题。作者同时指出，庭审中使用词典作为法律条款的解读依据有着很大的局限性，对词语"平义"的理解应该参考大规模语料库，同时考虑词语使用的语境和搭配问题。在这篇文章中，作者以"穆斯卡罗诉美国案"（Muscarello v. United States, 524 U.S. 125, 1998）为例，说明依赖词典理解法律条款可能带来的问题，并对基于语料库获取"平义"的方法进行系统论述，同时批评了法庭查询语料库时所使用方法的不当之处。在"穆斯卡罗诉美国案"这一案例中，穆斯卡罗涉嫌贩卖毒品。事发时他驾驶一辆卡车，卡车的手套箱内锁着一支手枪。法庭讨论的焦点问题仍在《美国法典》第18篇第一部分第924章（参见上文）的理解，但此次庭审讨论涉及的词是carry（携带）。穆斯卡罗认为，《美国法典》中的carry指的是"随身携带"，而法庭则认为，这里的carry应该包括"随车携带"，即"车载"。作者从大型语料库中提取了carry一词的所有搭配方式并逐个进行分析，以此证明carry的"平义"并不是"车载"。Mouritsen的文章不仅批评了词典方法的不可靠之处，还积极推广了基于语料库的方法。

自以上两篇文章发表后，美国法律从业者逐渐意识到语料库在庭审中可能起到的作用。由于语料库中收集了大量的语言使用实例，而且语料库中提供了词语使用的完整语境，更方便人们确定词语在语境中的"平义"。如今，利用大规模语料库分析关键词语的"平义"已经成为美国法庭解读法律条款的重要途径。

3. 美国法庭庭审案例三则

3.1 案例1："联邦通信委员会诉美国电话电报公司案"（FCC v. AT&T Inc., 562 U.S. 2011）

背景：2004年，美国电话电报公司（简称AT&T）和联邦通信委员会（简称FCC）达成一致，由美国政府投资，委托AT&T公司为全美学校和图书馆提供电信和因特网接入服务。同年8月，AT&T有关人员向FCC透露，公司存在收费过高行为。为此，FCC执法局（FCC Enforcement Bureau）展开了调查。此案以AT&T向政府退款50万美元而告结。然而，以CompTel公司为代表的部分消费者要求FCC向用户公开调查情况。对此，AT&T提出异议，认为其公司的隐私权受《信息自由法案》（*Freedom of Information Act*）的保护，FCC无权公开AT&T公司的隐私信息。根据《信息自由法案》第522条，公开此信息将"侵犯个人隐私"（invasion of personal privacy）。为了保护公司的隐私，AT&T向美国联邦上诉法院第三巡回审判庭（U. S. Court of Appeals for the Third Circuit）提出上诉，并于2009年9月22日得到该审判庭裁决的支持。审判庭认为，《信息自由法案》中的"个人"（personal）可以指代"公司"（corporation）。然而，FCC对这一审判结果表示不满，认为"个人隐私"只能指自然人的个人隐私，不适用于公司，并向美国最高法院申诉。2011年3月1日，最高法院撤销了第三巡回审判庭的裁决。法庭一致认定，AT&T公司的隐私不受《信息自由法案》的保护。

此案中的焦点问题是personal privacy（个人隐私）是否适用于公司。法官们认为，personal是person的形容词形式，但personal一词的意义并不取决于person。尽管person可以用来指代公司（即法人），personal却只能指代个人。法庭充分利用了语料库技术，从"当代美国英语语料库"中提取了personal与所有名词构成的各种搭配，如personal income, personal life, personal friend, personal fortune, personal experience, personal letter, personal attack, personal triumph, personal secretary, personal history, personal physician, personal popularity, personal problem, personal reason, personal relationship, personal use, personal loan, personal appearances, personal style, personal freedom, personal contact, personal appeal, personal choice, personal property, personal representative, personal interest, personal ambition, personal feeling等等，所有这些搭配中的personal都指"个人"，用personal指"公司"是史无前例的。

此案在美国法律界产生了很大的影响。从此，美国公司的隐私权不再受《信息自由法案》的保护。甚至有人认为，这是左翼势力试图控制美国企业的一大胜利。

3.2 案例2："美国诉科斯特洛案"（United States v. Costello, 666 F.3d 1040, 2012）

背景：伊利诺伊州一女孩，名字叫科斯特洛，住在距离圣路易斯不远的小城Cahokia。她与来自墨西哥的一非法入境男子同居近一年时间。2003年7月，该男子因贩卖毒品被捕入狱，刑满释放后被遣送回国。后来，该男子再次非法入境美国。2006年3月的某一天，

该男子到达圣路易斯某车站，科斯特洛驱车将他接到自己的住所，两人再次同居。2006年10月，该男子第二次因贩卖毒品、非法入境被警方拘捕，并被判重刑入狱。

根据《美国法典》第8篇第1324章第(a)(1)(A)(iii)条，任何人，“如果明知或无视外国人非法入境美国或非法居留美国，却将此外国人窝藏、包庇或庇护（或企图窝藏、包庇或庇护）于任何地方，包括任何建筑物或交通工具之中”（anyone who “knowing or in reckless disregard of the fact that an alien has come to, entered, or remains in the United States in violation of law, conceals, *harbors* or shields from detection [or attempts to do any of these things], such alien in any place, including any building or any means of transportation”），将处以5年以下监禁，并处25万美元以下罚款。

本案中的科斯特洛被指控触犯《美国法典》第8篇第1324章第(a)(1)(A)(iii)条中的“窝藏”（conceal）、“包庇”（harbor）和“庇护”（shield from detection）罪，数罪并罚。联邦地区法院法官认为，科斯特洛明知男子是非法入境者，且曾驱车到车站将男子接到自己的住所，有窝藏罪犯的企图，故判缓刑2年，并罚款200美元。后来，科斯特洛提出上诉，此案提交至美国联邦上诉法院第七巡回审判庭（U. S. Court of Appeals for the Seventh Circuit）。

美国联邦上诉法院第七巡回审判庭Posner法官认为，“窝藏”（conceal）指将犯人隐藏起来，不让外人发现，而“庇护”（shield from detection）指隔离起来，以防警方发现。从案件的基本事实看，没有证据能够证明科斯特洛犯有“窝藏”或“庇护”罪。因此，该案的关键是看科斯特洛是否犯有“包庇”（harbor）罪，这就要求我们弄清harbor这一词的常用意义。

Posner强烈反对通过词典来确定词汇的意义。他指出，“词典中的定义是脱离语境的，而句子的真正意义取决于语境，包括对语言产生背景的理解”。为了搞清harbor一词的意义，Posner使用了一种特殊的语料库——网络。随着互联网的不断普及和互联网数据的迅速递增，互联网上的文本数据已经成为世界上最大的语料库，从互联网上挖掘数据的方法也越来越多。西方有学者专门研究作为语料库的网络（Web as corpus）。Posner使用网络的方法十分简单。他在谷歌中搜索harbor的搭配词，结果发现，harbor之后最常见的搭配词包括fugitives（亡命者，50,800次）、Jews（犹太人，19,100次）、refugees（难民，4,820次）、enemies（敌人，4,730次）、Quakers（贵格会教徒，3,870次）等。这些数据清楚地表明，harbor的意思不是“为……提供住所”（这是词典中给的解释），而是“通过窝藏、转移到安全地点或提供人身保护等途径，有意识地保护某特定群体成员的安全，使其免受当局的伤害”。在Posner看来，本案中的科斯特洛只是想独自拥有男友，无意与当局为敌。

在审理此案的过程中，Posner不依赖传统的词典，而利用互联网上的大众语言作为证据，来解决法律文本中的“平义”问题，独具匠心。

由以上两则案例可以看出，作为科技发展和语言学研究结合的新兴事物，语料库已经成为美国法庭上的一种重要武器。这对我国的法制建设和驻外企业处理法律纠纷都具有重要启示。

3.3　案例3：Jared Diamond与巴布亚新几内亚

上文中提供的案例主要涉及的是美国法庭庭审过程中如何利用语料库来解决“平义原则”问题，此类案例还有“犹他州诉J.M.S案”（State of Utah v. J.M.S, 2011 UT 75）等。然而，语料库在美国法律中的应用并不仅限于此。下文介绍著名语料库语言学家Douglas Biber如何利用语料库语言学研究方法提供诉讼证据。

背景：美国著名科普作家、加州大学洛杉矶分校地理学教授、1999年度国家科学奖章获得者、普利策奖获得者Jared Diamond于2008年4月21日在《纽约人》（*The New Yorker*）杂志上发表一篇题为Vengeance is ours: What can tribal societies tell us about our need to get even?的文章，文章中描述了这样的故事。

Jared Diamond在巴布亚新几内亚（Papua New Guinea）作鸟类研究期间，结识了一位驾驶雪佛龙越野车的人，名为Hup Daniel Wemp（实质上，Wemp是世界野生动物基金会的工作人员）。在为Diamond当向导的过程中，Wemp向Diamond讲述了一系列故事，Diamond之后便将故事串联起来，撰文发表于《纽约人》杂志上。根据Diamond的描述，巴布亚新几内亚Handa部落的一头猪闯入了Ombal部落的花园中，引发了两部落间的冲突，在冲突中杀死了Wemp的叔叔。Wemp因此大为恼火，煽动了一项长达数年的复仇计划，1992年至1995年间，共召集士兵数百名，对Ombal部落发起了一轮又一轮的攻击，盗取Ombal部落300多头猪，先后杀死对方部落无辜人员29名，并花重金为士兵提供慰安妇。最终，Wemp得以复仇，并将一支利箭插入对方首领Henep Isum Mandingo的脊椎，致使Mandingo终生残疾。

Diamond的文章在《纽约人》杂志发表后，引发了人们广泛的关注。时隔一年，2009年4月，Wemp和Mandingo以诽谤罪一纸诉状，将Diamond和《纽约人》的出版商告上法庭，索赔10,000,000美元。控诉说，Diamond无中生有，故事中所讲的只有人名和部落名是真的，其他细节都是莫须有。因为Diamond的诽谤，Wemp已无法面对巴布亚新几内亚高地的乡亲父老。

至此，Diamond和Wemp各执一词，人们关注的焦点转到Diamond所写的故事是否属实呢？故事中所描述的人物及事件是否的确发生过？是Diamond在说谎还是Wemp不诚实？为此，人们经过多方调查，发现Wemp和Mandingo根本不相识，而且Mandingo并无残疾，种种议论变得对Diamond十分不利，人们开始试图揭穿Diamond的谎言，并为此付出不懈的努力。

受美国艺术科学研究实验室Rhonda Roland Shearer的委托，著名语料库语言学家Douglas Biber对Diamond故事的真伪进行了分析，并将分析结果公布于美国艺术科学研究实验室的网站http://www.imediaethics.org/。

在分析报告中，Biber以Linguistic analysis of quotations attributed to Daniel Wemp in New Yorker article为题，利用语料库语言学的研究方法，鉴别Diamond发表的故事中Wemp的话语是否的确是Wemp所说。在研究中，Biber通过量化手法，进行了一系列对比，认为Diamond所写故事中引用Wemp的所谓“口语”在词汇和语法方面并不具备口语特征，而

是Diamond编写的书面语，从而得出了Diamond编造故事的结论。例如，

a stone quarry *from which* the Ombal enemy was throwing stones

a night raid *in which* we sneak into an enemy village

据Biber的研究，以上两句是书面语中常见的表达方式，不具备口语特征，表明这些话很可能是Diamond编造出来的，而非Wemp亲口所说。

Biber以语料库语言学研究方法证明，Diamond编造故事，误导读者，从而为Wemp洗清了罪名，维持了公平，并为维护媒体道德作出了重要贡献。

我们希望国内法律界的研究者和从业人员也能借鉴美国同行的做法，为法律诉讼提供更为客观可靠的证据，同时希望语言学界的朋友们更多地考虑如何将语言学研究成果投入到应用之中。

4. 结语

模糊性是语言的本质特征之一，而与此相矛盾的是，法律语言需要精确无误，方可做到有法可依，因而对法律条文的正确解读成为法庭庭审中常常遇到的棘手问题，也是司法公正的前提。美国法律体系中为了保证“平义原则”得以落实，要求人们在解读法律条文时根据语境和搭配确定语义，这一点注定语料库在其中大有用武之地。

此外，近年来语料库语言学研究方法的发展使得语料库能够更好地为法庭诉讼提供帮助，可以有效地为诉讼提供证据，这必将导致语料库语言学与法律语言学的结合。

语料库在法律领域的应用是语料库应用的又一范例。大规模语料库的建成和语料库语言学的发展应该引起国内法律研究者、司法从业人员和语言学研究者的重视。

参考文献

Apperson, G. 1993. *The Wordsworth Dictionary of Proverbs* [M]. Ware: Wordsworth Editions.

Biber, D. 2009. Linguistic analysis of quotations attributed to Daniel Wemp in *New Yorker* article. Art Science Research Laboratory’s Research Project [OL]. http://www.imediaethics.org/content/II6_report.pdf (accessed 4/6/2014).

Cunningham, C. & C. Fillmore. 1995. Using common sense: A linguistic perspective on judicial interpretations of “use a firearm” [J]. *Washington University Law Quarterly* 73(3): 1159-1214.

Firth, J. 1957a. A synopsis of linguistic theory 1930–55 [A]. In F. Palmer (ed.). 1968. *Selected Papers of J. R. Firth 1952–1957* [C]. London: Longmans. 168-205.

Firth, J. 1957b. Modes of meaning [A]. In J. Firth. (ed.). 1957. *Papers in Linguistics 1934–1951* [C]. London: OUP. 190-215.

Mouritsen, S. 2011. The dictionary is not a fortress: Definitional fallacies and a corpus-based approach to plain meaning [J]. *Brigham Young University Law Review* (5): 1915-1980.

Sinclair, J. 2008. The phrase, the whole phrase and nothing but the phrase [A]. In S. Granger &

F. Meunier (eds.). *Phraseology: An Interdisciplinary Perspective* [C]. Amsterdam: John Benjamins. 407-410.

Whitsitt, S. 2005. A critique of the concept of semantic prosody [J]. *International Journal of Corpus Linguistics* 10(3): 283-305.

梁茂成，2012，语料库语言学研究的两种范式：渊源、分歧及前景 [J]，《外语教学与研究》（3）：323-335。

梁茂成，2014，语料库语言学的核心议题及学科前景 [J]，《解放军外国语学院学报》（1）：5-7。

王国轩、王秀梅，2009，《孔子家语》[M]。北京：中华书局。

通信地址：100089 北京市北京外国语大学中国外语教育研究中心

四大名著汉英平行语料库的宏观语言特征研究*

南开大学/燕山大学 刘泽权 燕山大学 刘鼎甲

提要：本文利用已建成的“四大名著汉英平行语料库”，考察其11个英文全译本在词汇多样性（包括标准类/形比与名词/代词比）、词长分布、词汇密度、平均句长以及汉英句对应类型等多个方面的特征，并在此基础上对比了各译本翻译策略与风格的异同及成因。研究发现，11个英译本在很多方面存在较高比例的一致性，表现出显化特性，各译本作为翻译语言和小说文体的特征明显，但又呈现出各自独特的风格。本文以多译本真实、客观的数据进一步验证了语料库方法对于客观描述译者风格的意义。

关键词：四大名著、翻译、平行语料库、数据分析、译者风格

1. 引言

中国古典文学四大名著（即《三国演义》、《水浒传》、《西游记》、《红楼梦》，以下简称“四大名著”）一直是国内文学、语言学、译界研究的焦点之一。“在西方，几个世纪以来对中国古典小说的翻译，特别是与俗文学有关的章回小说的翻译，持续不断，经久不衰”（汪榕培、王宏 2009：209）。自19世纪末至今，四大名著的数十个全译本、节译本在国内外接连问世（详见刘泽权、谭晓平 2010：81-86）。这些译本对于我国文学典籍的英译、传播、翻译研究乃至翻译教学具有巨大的现实意义与参考价值。然而，四大名著英译本的研究在宽度和深度上仍存在着较大的局限性，大部分研究从篇章角度探讨具体文化现象、译者局部翻译策略、修辞手法等，是典型的基于“点”的“共时”研究，对于多译本主体风格、篇章特点、翻译策略的比较等基于“面”的“历时”研究比较罕见，且定性分析居多。具体说来，国内对《红楼梦》英译本译者风格的研究业已展开，亦取得了一些实质性进展，但以典籍英译多文本，特别是四大名著多译本为对象的定量研究较少。究其根由，主要是因为四大名著内容庞大、汉语版本繁杂，特别是以词语、句子为翻译单位的翻译对比研究工作量巨大，难以开展彻底、全面的分析和比较。因此，建设四大名著平行语料库并开展基于语料库的翻译研究尤为重要。目前，该库已建设完毕并实现《三国演义》、《水浒传》、《西游记》和《红楼梦》（以下分别简称《三》、《水》、

* 本文为国家社科基金项目“《红楼梦》平行语料库中的汉英文化词典编纂研究”（10BYY011）、河北省社科基金项目“中国古典文学四大名著汉英平行语料库及检索平台的创建”（HB08BYY008）及“基于语料库的翻译教学模式研究”（HB10Y007）的阶段性成果。

《西》、《红》）及其多个全译本的电子化以及汉英一语多译文本的句级对齐，不仅方便了汉语典籍及其英译的保存，也使得研究者能够针对各译本开展系统、全面的研究，亦可视“四大名著”翻译为我国典籍小说英译的一段历史长河，去探究我国小说外译传播及其与英美主流文学碰撞及被接受的历程和经验。本文拟从词汇多样性（包括标准类/形比，名词/代词比）、词长分布、词汇密度、句长分布以及汉英句对应类型等多个方面对四大名著各译本进行对比分析，试图发现各译本的风格与翻译策略的异同及成因，为大规模的典籍英译研究提供初步参考。

2. 语料库与译者风格研究

近年来，随着翻译学研究的不断深入，“译者风格”作为一个独立的研究内容逐渐进入研究视线。不少学者发现，大多数译文不仅体现源文风格，同时也存在“译者的声音”（translator’s voice）（Hermans 1996：27）。House（1977，1997）认为译者风格是衡量译文语言“正式程度”的一个变量。Baker（2000）将译者风格定义为一系列语言或非语言特征所表现出来的个性特征。张美芳（2002：57）则指出，通过有别于其他译者的个性化语言习惯，译者在译本中留下了自己的翻译风格，具体表现在文本类型和翻译策略的选择以及译者所运用的前言、后记、脚注、文内注释等释义方法。

为探寻“译者风格”的根源，国内学者在多个方面展开了研究。赵巍、孙迎春（2004）从“个人方言”对翻译过程中理解与表达的影响入手，揭示了译者风格存在的必然性。姚琴（2009）对比了《红》的霍克思译本和杨宪益译本中“文字游戏”的翻译，发现霍、杨译本分别呈现归化、异化的风格。上述研究以经验、内省型分析为主，对文体、风格的论述有限，特别是“在于对长篇译作的文体风格进行量化研究时，可操作性往往会成为研究中的一大难题（肖维青 2009：254）”。因此，将语料库的方法引入译者风格的考察非常必要。

自Baker（1993）倡导将语料库应用于翻译研究以来，语料库语言学方法已经被广泛应用于译者风格考察、翻译普遍性验证、翻译教学以及词典编纂等理论与实践的诸多方面并取得了丰硕成果。Olohan & Baker（2000）利用“英语翻译语料库”（TEC）文学子库与英国国家语料库（BNC）的文学子库对say/tell + that结构进行了量化研究。Kamenická（2008）利用平行语料库探讨了散文的翻译显化现象及其对于译者风格的影响。在国内，冯庆华（2008）通过母语文化视角分别从修辞、文化词汇、英语习语、词频等方面对《红》的霍译本风格进行了立体式的审视和评介，发现霍译本语言生动、词汇量大、书面语口语风格界限分明以及归化处理等特点。王克非（2003，2004）应用通用汉英平行语料库考察英汉、汉英句对应现象，发现无论英译汉或汉译英都呈现目标语文本扩增特点。此外，一些研究依据Baker的研究方法，分析译语的语言特征，如：王峰、刘雪芹（2012）以《木兰辞》英译文为例，黄立波、朱志瑜（2012）以葛浩文的中国现当代小说英译为例，卢静（2013）以《聊斋志异》译本为例，分别以源语文本和译语文本为基点对译者风格展开了讨论。刘泽权等（2011）对《红》的三个英译本分别从类符/形符比、词汇密度、词长分布、平均句长和汉英句子对齐类型等方面全面地对比分析了译者的不同风格。刘

泽权、朱虹（2008）、刘泽权、闫继苗（2010）、刘泽权、田璐（2009）、刘泽权、刘艳红（2011）还从《红》的习语、叙事标记语及报道动词使用及其在四个英译本中的翻译等微观方面进行了系统考察与分析。

上述研究覆盖面广，分析深彻，部分研究还将译文语料库与英语本族语言者语料库进行对比，所得结果极具参考价值。遗憾的是，大多数研究未对同源多译本进行语内对照考察，因而无法对译者风格进行总体的把握。其次，一些研究过于主观，对于译者风格与作者风格对译者造成的影响未做明确区分。此外，针对《红》译本的研究较多，鲜有针对《西》、《水》、《三》多译本的译者风格考察。作为尝试，本文对四大名著11个英文全译本的语言特征进行文本间可验证性的分析，试图揭示各个译本的译者风格，以期从历时和共时、横向和纵向多维度全面斑窥四大名著英译的整体风貌。

3. 译本数据统计与分析

3.1“四大名著”汉英平行语料库概况

四大名著汉英语料库包括《西》的余译、詹译100回全译本，《三》的罗译、邓译120回全译本，《水》的沙译100回全译本、敦译的120回全译本及赛译的70回全译本，《红》的乔译前56回全译本，邦译、霍译及杨译的120回全译本，汉英文本的具体数据见表1。

表1. “四大名著汉英语料库”基本数据

文本名称	文本类别	文本信息		文本数据	
		作者/译者	出版年代	章回	形符
《西游记》	源文	吴承恩	2003	100	715,367
	译文	Anthony Yu 余国藩	1977	100	701,087
		W. J. F. Jenner 詹纳尔	1985	100	652,198
《三国演义》	源文	罗贯中	2000	120	894,273
	译文	Moss Roberts 罗慕士	1976	120	543,939
		C. H. Brewiit-Taylor 邓罗	1925	120	576,084
《水浒传》	源文	施耐庵	2006	70	828,141
	译文	Sidney Shapiro 沙博理	1980	100	534,192
		John & Alex Dent-Young 敦特·杨父子	1994	120	737,729
		Pearl Buck 赛珍珠	1933	70	565,521

（待续）

（续表）

文本名称	文本类别	文本信息		文本数据	
		作者/译者	出版年代	章回	形符
《红楼梦》	源文	曹雪芹	2007	120	875,818
	译文	Bencroft Joly乔利	1892–3	56	441,939
		R.S. Bonsall邦斯尔	1948	120	838,159
		David Hawkes & John Minford霍克斯、闵福德	1973–1986	120	830,443
		Yang Hsien-yi & Gladys Yang杨宪益、戴乃迭	1978	120	625,988

3.2　译本数据统计与分析

本文分别使用笔者自己编写的Segmentator 1.2软件对文本进行分句和分词，利用自然语言工具NLTK 3.0（Steven *et al*. 2009）对所有语料进行了词性标记，并借助语料库工具WordSmith Tools 5.0统计出四大名著各译本的类符/形符比、标准化类符/形符比、名词/代词比，词长分布、高频词、词汇密度、句数及平均句长，结果见表2、3、4。标准类符/形符比、名词/代词比能够综合反映文本的词汇丰富程度和多样性。Semenza, Mondini & Borgo（2003）指出，较之名词，代词对事物的称呼需要依据语境来判断，因而大量使用代词会使得文本相对更加“模糊”，而名词的大量使用则表现出文本更加“明晰”的特性。其他有关考察内容的概念、统计方法与意义，见刘泽权（2010）及刘泽权等（2011）。

表2. “四大名著”各译本参数统计

考察项	《西游记》		《三国演义》		《水浒传》			《红楼梦》			
	余译	詹译	罗译	邓译	沙译	敦译	赛译	乔译	邦译	霍译	杨译
形符	701,087	652,198	543,939	576,084	378,729	468,893	565,521	441,939	838,159	830,443	625,988
类符	16,462	14,864	16,673	14,661	12,080	14,297	8,623	14,113	16,277	23,865	17,486
类/形比	2.35	2.28	3.07	2.54	3.19	3.05	1.52	3.19	1.94	2.87	2.79
标准类/形比	43.18	42.07	45.42	42.17	43.97	42.3	34.45	43.83	37.30	42.78	43.75
名词/代词比	2.03	1.97	2.93	2.10	2.02	1.91	1.69	1.82	1.51	1.68	1.60
平均词长	4.29	4.24	4.39	4.1	4.27	4.22	3.91	4.31	4.17	4.32	4.28

3.2.1　词汇层面

在形符方面,《西》的余译略微高于詹译;《三》的邓译较罗译略高;《水》的赛译最多，高出沙译约49%，敦译居中;《红》的邦译最高，霍译次之，二者均高于杨译约34%之多。由于目的语扩增是翻译显化现象的特征之一（Mauranen 2005：93-100；刘泽权、侯羽 2008：55-58），因此我们可以推断余译的显化程度略高于詹译，邓译的显化程度略高于罗译，赛译的显化程度最为明显，而沙译的显化程度最不明显，邦译的显化程度较高，霍译次之，杨译最低。

从表2可以看出，在类符数量方面,《西》的余译较之詹译要多,《三》的罗译较邓译多,《水》的敦译最多、沙译次之、赛译最少,《红》的霍译最多、杨译次之、邦译最少。这在一定程度上反映了余译、罗译、敦译和霍译词汇丰富，可能预示着这四个译本在叙事和人物描写方面或更为生动，用词更加准确具体。标准类符/形符比的统计结果也基本印证了这一观察:《西》的余译略高于詹译,《三》的罗译明显高于邓译,《水》的沙译略高于敦译并大幅度领先于赛译,《红》的乔译、杨译略高于霍译并远远超过邦译。尤其是《水》的沙译和《红》的杨译，似乎严格遵循原文的叙事结构，既降低了译文长度，同时再现了原文的人物情节描写。需要指出的是，除《水》的赛译和《红》的邦译外，其他九个译本的标准类/形比值均接近TEC中小说类子库的标准类/形比44.63（Olohan 2004：80），体现了明显的译文特征。赛译、邦译的标准类/形比值与各自同一源语译本存在较大的差距，但在形符数量上远远领先，可见赛译、邦译在翻译中一方面尽可能在文内进行释义说明（属显性翻译），导致文本长度扩增，另一方面有意识地减少词的类符数目，降低译文阅读难度。

名词/代词比例的考察表明，类/形比越高的译本，其名词/代词比例亦越高，该结果与类/形比的考察结果具有高度相关性，亦表明文本的词汇丰富程度越高，译者也更偏向变换词汇来表达文本意义。表2显示,《西》的余译略微高于詹译,《三》的罗译高于邓译；受总体形符数的影响,《水》的赛译最低、敦译居中、沙译最高,《红》的乔译最高、霍译次之、邦译最少。因此可以推测，余译的显化程度略高于詹译，罗译的显化程度略高于邓译，乔译的显化程度较高，而杨译和邦译的显化程度最低。

词长方面，除了《水》的赛译为3.91，其余10个译本平均词长均在4.1-4.4之间上下变动，趋近于TEC平均词长4.36（Olohan 2004：80），呈现出翻译的普遍性特征。其中,《西》的余译略高于詹译,《三》的罗译较之邓译大幅度增高,《水》的沙译最高,《红》的霍译最高，乔译、杨译次之，邦译最短。这与上文各译本标准类/形比的比较结果不谋而合，也从另一个方面反映了译者们的词汇丰富程度差异。各译本中，3字母词的数量最大，其次是4字母词和2字母词。将词长统计结果绘制成点线图可使得观察和对比更加直观，详见图1。

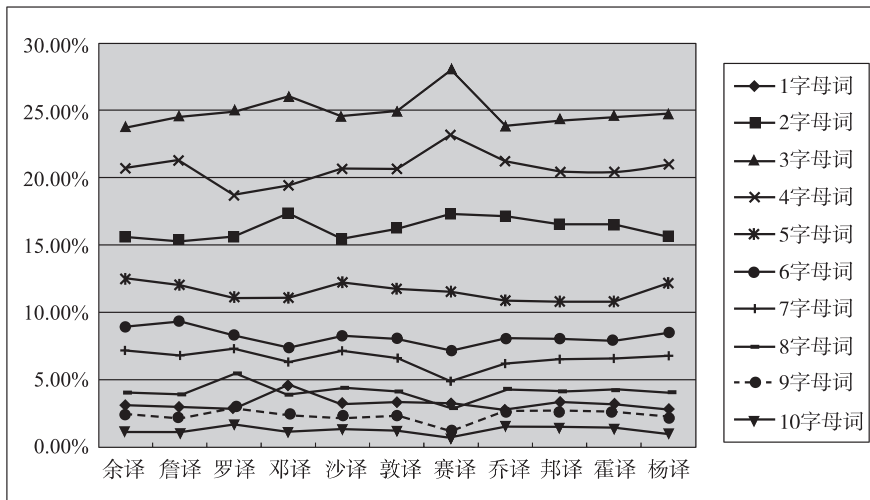


图1.　四大名著各译本词长统计分布图

具体说来，11个译本中的3字母词约占全文25%左右，其他高频词的长度依次是4字母词（20%左右）、2字母词（15%以上）、5字母词（10%以上）和6、7字母词（均在5%以上）。上述结果与刘泽权（2010：80）针对《红》三个英译本词长分布的考察结果基本一致。译文库中，较短单词（2、3、4字母单词）约占译文总字数的60%，值得关注。与词频的考察对照可知，这些短词多为虚词和人称代词，这亦可充分说明短词的大量使用是译文口语化和小说性的特征。其中，《西》的詹译本所使用的3、4字母词多于余译，而在5—9字母及以上的词略少于余译本；《三》的邓译本1—5字母词的比率均高于罗译本，而在6字母及以上词的使用上少于罗译；《水》的赛译本所使用的2、3、4字母词均多于其他两个译本，其次是敦译、沙译，而在6字母及以上单词的使用方面，沙译的比例最大，敦译居中；《红》的乔译、邦译和霍译在词汇长度的使用方面差别不大，杨译所使用的2、3、4个字母词都少于其它三个译本，而使用5、6、7个字母词则多于其他三个译本。由于大量使用长度较短的单词可降低阅读的难度，因此可以推测《西》的詹译、《三》的邓译、《水》的赛译及《红》的邦译更易阅读。

本文也对各译本中前24个高频词进行了统计（见附表1）。结果显示，除赛译《水》外，其他10个译本中使用频率最高的五个词同为虚词 the、and、to、of、a 且排序基本相同，这一发现与Olohan（2004）对TEC的考察结果完全一致，而与BNC的排列顺序出入较大，这也体现了文学译语的共性特征。此外，11个英译本中处在前24位的高频词多为无实际意义的功能词，如冠词（the、a）、介词（of、with、for）、系动词（is/was）、连词（and、but、as）以及人称代词（you、I、she/her、he）。作为功能词，人称代词一般用于显式地指称话语发出者、写作者、话语的接受者与可作明显区分的事物或人物。Biber（1999：328-334）发现，一般情况下，在口语和小说语体中，为明确指称参与者、表达思想与行为、展现人物对话，人称代词的使用频率较高（分别为13.1%和9%），而在新闻和学术语体中，由于顾及语体的正式性和客观性，人称代词的使用频率较低（分别为3.1%和1.9%）。在

11个译本中，各译本人称代词所占比率分别为：余译10.02%、詹译10.24%、罗译8.63%、邓译10.03%、沙译9.74%、敦译9.87%、赛译11.59%、乔译10.35%、邦译11.30%、霍译11.25%、杨译11.30%。除罗译本外，其他各译本均高于9%，这与Biber的发现基本一致，进一步说明译语明显属于小说类文体。此外，各译本在人称代词的使用上呈现出不同的特点：各个译本人称代词的使用基本相似，You、I所占比率颇高（分别在1.2%和0.84%以上），He、She、It、They等第三人称代词频率亦较高。人称代词使用的相似性再现了原文叙事与对话兼容并蓄的语体特征：《三》的两个译本中，I、You等人称代词频率较之其他各个译本低且差距较大，而He、Cao、Ts'ao等代词、专有名词频率非常高，邓译本更是达到了1.96%，明显地呈现出《三》以叙事为主的特点。《红》的各个译本中人称代词She和Her（平均约1.2%）的比率与其他七译本形成强烈的反差（约0.1%），这说明《红》中女性角色较之其他作品要多，因而占主导地位。

除此之外，报道动词said在各个译本的出现频率也非常之高。Baker（2000）对TEC中报道动词say的统计分析证实报道动词的选择可以反映译者风格。刘泽权（2010：88）认为，人物话语是小说的重要组成部分，译者在译文中大量使用报道动词，明示原文中隐含的叙述方式，具备显化特征。通过对翻译策略和技巧的分析可以发现："说（道）"类报道动词在四大名著原文中的使用单一而且频繁，译文中say的使用越多可说明译者采取直译策略倾向越明显（同上）。附表1显示，对于报道动词say的应用，《西》的余译（1.02%）高于詹译（0.77%），《三》的邓译（0.68%）高于罗译（0.58%），《水》的赛译（0.79%）和敦译（0.78%）均高于沙译（0.53%），《红》的邦译（1.19%）远高于霍译（0.70%），杨译（0.34%）和乔译（0.16%）较低。可以看出，余译、邦译更倾向使用say作报道动词并采用直译的对话策略，而乔译、杨译、沙译和罗译更倾向于使用其他报道动词和非对话式的叙事方式。其次，从译者风格上看，say出现的位置与译者的叙事风格有直接关系：（1）如果译者采用对话体的直接引语来再现原文的人物对话和叙事方式，多倾向于使用[直接引语+人名（人称代词）+ said]的叙事方式；（2）如果译者选择间接引语的叙事方式，则会更偏向于将原文的人物对话及叙事套语合并为[人名（人称代词）said +（that）+间接引语]的叙事方式。这一推论与下文关于that的考察结果一致。

值得注意的是，在11译本中，that有不同程度的呈现。Biber（1999：351，674）指出，that在文本中反复出现一般有三种功用，即作为指示代词、指示限定词兼谓语补语子句的引导词。在口语与小说文体中，that作为指示代词和指示限定词出现的几率分别为60%和30%，远超较为正式的学术与新闻语体。that作为谓语补语子句引导词在口语与小说文体中出现的频率分别为0.71%和0.58%，而在新闻和学术文体中分别为0.52%和0.21%。that用于引导间接报道引语时受动词的支配，借此来转述小说人物间的话语交流。可见，that的频率越高，则语体越不正式，显化程度就越高。附表1显示，11个译本中that的频次均在0.53%以上，且多处在0.8%–1%之间，尤其是在《红》的四个译本中多在1%以上，这充分表明11个译本作为小说的文体特征明显。

词汇密度方面，表3显示，各个译本的差别较大，但都处在40–50之间，尤其是詹译、邓译和赛译，低于Laviosa（1998）对英语原语语料库的考察结果（54.95），可见各译本作为文学译文的语言特征比较明显，这一发现与标准类/形比的考察结果一致。只有罗译的

词汇密度较接近英语原创小说，杨译、霍译再次之。

表3. “四大名著”各译本词汇密度统计

考察项＼译本	《西游记》		《三国演义》		《水浒传》			《红楼梦》			
	余译	詹译	罗译	邓译	沙译	敦译	赛译	乔译	邦译	霍译	杨译
实词	316,868	285,997	281,994	251,727	178,720	206,023	230,701	210,378	396,751	408,367	312,911
形符	701,087	652,198	543,939	576,084	378,729	468,893	565,521	441,939	838,159	830,443	625,988
词汇密度	45.20	43.85	51.84	43.70	47.19	43.94	40.79	47.60	47.34	49.17	49.99

3.2.2　句子层面

为考察11个译本的平均句长，我们使用Word对各译本进行了统计（见表4）。Laviosa（1998）指出，翻译的叙述文体的句子明显长于原文文本，其对英语译语及原语语料库平均句长的考察结果分别为24.1和15.6。表4显示，除《水》的赛译和《红》的乔译外，其余九个译本平均句长均低于15.6，沙译甚至低至10.9。柯飞（2005）指出，翻译中的显与隐是共存的，隐显现象及其程度可能与语言的形式化程度和翻译方向相关。汉语是意合语言，其形式化程度远低于英语，加之四大名著半文半白的语言特性，汉语句子通常较英语要短，且多为主动句，译文受源语这一特点的影响，平均句长多处于14—15个形符之间。从表4可观察到，赛译和乔译超出20个形符，与在高频词考察中所发现的两译本较多使用and、that等连词的情况相互印证。尤其是赛译，其and的频率位于该译本前24位高频词之首，达到了6.17%。由于赛译成稿于1933年，乔译成稿于1892年，明显早于其他译本，这可能是二者较多使用长句的原因之一，反映出早期的英语译语语体较为正式，表现出与现代英语的历时性区别。这一推断也与刘泽权等（2011：62）针对《红》的乔译本平均句长的观察不谋而合。

表4. “四大名著”各译本句子信息统计

考察项＼译本	《西游记》		《三国演义》		《水浒传》			《红楼梦》			
	余译	詹译	罗译	邓译	沙译	敦译	赛译	乔译	邦译	霍译	杨译
句子数目	43,221	38,580	37,549	34,165	34,200	30,951	25,958	21,362	59,936	54,068	45,898
平均句长	14.7	15.4	14	15.5	10.9	14.7	21.3	20.69	13.98	15.36	13.64

3.3.3　汉英句对应类型

通过对句子层面对齐的四大名著汉英平行语料库进行统计，可以得出汉英句子的对齐类型结果。从表5可以看出，各译本中一对一的汉英句对类型最多，占所有翻译句对类

型的三分之一以上，有六个译本超过了50%，个别达到了59%，说明译者们翻译的出发点基本是以句子为基本单位、以句句对应为再现策略。这里有两个现象值得关注：首先，11个译本中，《红》的四译本在这方面表现尤为显著，它们的一一对应类型均在53%以上，一二对应均在22%以上，两者之和接近80%，表明译文受源语文本叙事方式和结构的影响较重，对话型内容较其他七译本大，而叙事性内容较之其他译本要少，因此译本的"翻译腔"较为明显，这与《西》这种叙事性内容较多的文本形成了鲜明对比。其次，在二对一、二对二的对应类型中，《水》的赛译本和《西》、《三》的译本的运用比例基本在25%以上，赛译本甚至达到了35%以上，远远高于其他译本，可以推测这些译文试图调整译文的叙事结构，从而导致译文句子的合并和扩增。

表5. "四大名著"各译本汉英句对应类型

对应类型	《西游记》		《三国演义》		《水浒传》			《红楼梦》			
	余译%	詹译%	罗译%	邓译%	沙译%	敦译%	赛译%	乔译%	邦译%	霍译%	杨译%
1—1	31.00	34.24	46.75	48.75	55.32	59.18	39.72	56.11	56.83	53.64	59.23
1—2	24.27	21.80	13.46	11.72	19.93	16.62	10.76	23.36	25.62	25.34	22.79
1—3	4.22	3.09	2.52	2.41	5.01	3.57	2.27	5.31	8.77	7.46	3.57
1—4	2.11	1.59	0.78	0.99	1.20	0.95	0.44	0.97	2.79	1.88	0.53
1—5+	0.77	0.52	0.35	0.39	0.35	0.28	0.21	0.37	1.43	0.67	0.16
2—1	14.62	16.57	16.20	16.11	6.05	8.35	25.83	7.48	1.21	5.41	7.53
2—2	11.87	11.30	10.39	8.65	6.99	6.31	9.62	3.00	0.78	2.65	3.36
2—3	4.25	2.90	2.91	2.49	2.72	2.00	2.30	0.94	0.38	0.97	0.77
2—4+	1.77	1.00	0.96	0.91	1.06	0.73	0.77	0.39	0.31	0.45	0.22
3—1	0.91	2.76	1.93	2.66	0.29	0.62	3.18	1.01	0.06	0.62	0.94
3—2	1.27	1.56	1.65	1.98	0.41	0.60	2.14	0.37	0.05	0.37	0.40

4. 结语

本文从译本的语言特征入手，对四大名著的11个英译本进行了分析和探讨，所得数据为基于语料库的多译本译者风格考察提供了可靠依据。但是，由于考察项目的外在性所限，加上译者受源文与各自所处时代、社会及个人等诸多因素的影响，研究尚未深入，尤其是基于语境的考察有待展开。区分S-型和T-型风格（即译语的语言风格是受源语文本影响还是受译者风格影响）的研究亟需同时结合译文内容及对语境的分析研判，方可得出定量和定性相结合的结论。再次，本文的考察仅关注古典章回体小说英译文的译者风格，对

于源语文本的语体对译者风格的影响未作全面考察，因此，以语体为参考变量的译者风格对比尚需考证。最后，现有的译者风格研究都是通过对某个译本或者某几个译本进行的小范围尝试，所以难以区分所谓的“译者风格”是见诸所有翻译的“共性”，还是某个译者的“个人语言选择”偏好。因此，如何为“翻译普遍性”和“译者风格”建立相互区别的语言学模型是一个亟待解决的问题。

参考文献

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications [A]. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.). *Text and Technology: In Honour of John Sinclair* [C]. Philadelphia: John Benjamins. 233-250.

Baker, M. 2000. Towards a methodology for investigating the style of a literary translator [J]. *Target* 12(2): 241-266.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Spoken and Written English* [M]. New York: Longman.

Hermans, T. 1996. The translator's voice in translated narrative [J]. *Target* 8(1): 23-48.

House, J. 1977. *A Model for Translation Quality Assessment* [M]. Tübingen: TBL-Verlag Narr.

House, J. 1997. *Translation Quality Assessment: A Model Revisited* [M]. Tübingen: G. Narr.

Kamenická, R. 2008. Explicitation profile and translator style [A]. In A. Pym & A. Perekrestenko (eds.). *Translation Research Projects* [C]. Tarragona: Universitat Rovira Virgili. 117-130.

Laviosa, S. 1998. Universals of translation [A]. In M. Baker (ed.). *Routledge Encyclopedia of Translation Studies* [M]. London: Routledge. 557-570.

Mauranen, A. 2005. Translation universals [A]. In K. Brown (ed.). *Encyclopedia of Language and Linguistics* [M]. Boston: Elsevier.

Olohan, M. 2004. *Introducing Corpora in Translation Studies* [M]. New York: Routledge.

Olohan, M. & M. Baker. 2000. Reporting *that* in translated English: Evidence for subconscious processes of explicitation [J]. *Across Language and Cultures* 1(2): 141-158.

Semenza, C., S. Mondini & F. Borgo. 2003. Proper names in patients with early Alzheimer's disease [J]. *Neurocase* 9(1): 63-69.

Steven, B., E. Klein & E. Loper. 2009. *Natural Language Processing with Python* [M]. California: O'Reilly Media.

冯庆华，2008，《母语文化下的译者风格——〈红楼梦〉霍克斯与闵福德译本研究》[M]。上海：上海外语教育出版社。

黄立波、朱志瑜，2012，译者风格的语料库考察——以葛浩文英译现当代中国小说为例 [J]，《外语研究》（5）：64-71。

柯　飞，2005，翻译中的隐和显 [J]，《外语教学与研究》（4）：303-307。

刘泽权，2010，《〈红楼梦〉中英文语料库的创建及应用研究》[M]。北京：光明日报出版社。

刘泽权、侯　羽，2008，国内外显化研究现状概述 [J]，《中国翻译》（5）：55-58。

刘泽权、刘超朋、朱　虹，2011，《红楼梦》四个英译本的译者风格初探 [J]，《中国翻译》（1）：60-64。

刘泽权、刘艳红，2011，初识庐山真面目——邦斯尔英译《红楼梦》研究（之一）[J],《红楼梦学刊》（4）：30-52。

刘泽权、谭晓平，2010，面向汉英平行语料库建设的四大名著中文底本研究 [J],《河北大学学报（哲学社会科学版）》（1）：81-86。

刘泽权、田　璐，2009,《红楼梦》叙事标记语及其英译——基于语料库的对比分析 [J],《外语学刊》（1）：106-110。

刘泽权、闫继苗，2010，基于语料库的译者风格与翻译策略研究——以《红楼梦》中报道动词及英译为例 [J],《解放军外国语学院学报》（4）：87-92。

刘泽权、朱　虹，2008,《红楼梦》中的习语及其翻译研究 [J],《外语教学与研究》（6）：460-466。

卢　静，2013，基于语料库的译者风格综合研究模式探索——以《聊斋志异》译本为例 [J],《外语电化教学》（2）：53-58。

王　峰、刘雪芹，2012，基于语料库的译者风格研究：以《木兰辞》译文为例 [J],《广西民族大学学报（哲学社会科学版）》（2）：182-188。

王克非，2003，英汉/汉英语句对应的语料库考察 [J],《外语教学与研究》（6）：410-417。

王克非，2004,《双语对应语料库：研制与应用》[M]。北京：外语教学与研究出版社。

汪榕培、王　宏，2009,《中国典籍英译》[M]。上海：上海外语教育出版社。

肖维青，2009，语料库在《红楼梦》译者风格研究中的应用——兼评《母语文化下的译者风格——〈红楼梦〉霍克斯与闵福德译本研究》[J],《红楼梦学刊》（6）：251-261。

姚　琴，2009,《红楼梦》文字游戏的翻译与译者风格——对比Hawkes译本和杨宪益译本所得启示 [J],《外语与外语教学》（12）：50-52。

张美芳，2002，利用语料库调查译者的文体——贝克研究新法评介 [J],《解放军外国语学院学报》（3）：54-57。

赵　巍、孙迎春，2004，个人方言与文学翻译中的译者风格 [J],《外语教学》（3）：64-68。

通信地址： 300071 天津市南开大学外国语学院/066004 河北省秦皇岛市燕山大学出版社（刘泽权）

066004 河北省秦皇岛市燕山大学信息科学与工程学院（刘鼎甲）

附表 1. “四大名著”各译本高频词统计（%）

译本	《西游记》		《三国演义》		《水浒传》			《红楼梦》			
序号	余译	詹译	罗译	邓译	沙译	敦译	赛译	乔译	邦译	霍译	杨译
1	THE/6.95	THE/7.16	THE/6.02	THE/6.18	THE/6.09	THE/6.45	AND/6.17	THE/4.44	THE/4.23	THE/4.44	THE/4.29
2	AND/3.18	AND/3.45	AND/3.43	AND/4	AND/3.55	AND/3.64	THE/5.74	AND/3.53	AND/3.76	TO/3.47	TO/3.46
3	TO/2.99	TO/2.96	TO/3.31	TO/3.27	TO/2.82	TO/3.28	TO/2.76	TO/3.26	TO/3.04	AND/3.03	AND/2.86
4	OF/2.35	OF/2.07	OF/2.15	OF/2.57	A/2.16	A/2.09	HE/2.25	OF/2.68	OF/1.91	OF/2.37	A/1.82
5	A/1.76	A/1.95	A/1.57	HE/1.96	OF/1.83	OF/1.95	OF/2.05	A/1.92	A/1.81	A/2.03	OF/1.73
6	YOU/1.46	YOU/1.35	HIS/1.53	A/1.7	HE/1.6	HE/1.5	A/1.55	IN/1.6	SHE/1.52	YOU/1.46	YOU/1.43
7	HE/1.46	HE/1.27	IN/1.25	HIS/1.63	IN/1.36	YOU/1.42	I/1.32	THAT/1.3	THAT/1.43	IN/1.41	HER/1.39
8	HIS/1.07	IN/1.27	HE/1.24	IN/1.25	YOU/1.36	IN/1.31	IN/1.23	SHE/1.23	IN/1.41	HER/1.41	IN/1.3
9	IN/1.06	HIS/1.2	CAO/1.16	WAS/1.18	HIS/1.24	HIS/1.1	YOU/1.21	YOU/1.2	YOU/1.38	THAT/1.2	SHE/1.27
10	SAID/1.02	WAS/0.93	I/0.84	I/1.08	I/0.97	I/1.06	HIS/1.16	HER/1.16	IT/1.37	SHE/1.17	HE/1.02
11	THAT/0.98	I/0.92	YOU/0.82	YOU/1.05	WAS/0.94	WAS/1.05	IT/1.02	HE/0.95	WAS/1.27	I/1.14	THAT/0.97
12	I/0.87	THAT/0.89	WITH/0.72	THAT/0.76	WITH/0.82	WITH/0.84	THEY/0.94	AS/0.95	HE/1.22	WAS/1.14	I/0.94
13	WITH/0.72	WITH/0.87	FOR/0.69	HIM/0.76	HIM/0.74	IT/0.81	WAS/0.91	WITH/0.94	SAID/1.19	IT/1.12	WAS/0.92

（待续）

（续表）

译本	《西游记》		《三国演义》		《水浒传》			《红楼梦》			
序号	余译	詹译	罗译	邓译	沙译	敦译	赛译	乔译	邦译	霍译	杨译
14	WAS/0.69	IT/0.77	WAS/0.66	IS/0.76	ON/0.71	SAID/0.78	THEN/0.9	IT/0.93	I/1.13	FOR/0.92	FOR/0.9
15	FOR/0.67	SAID/0.77	HIM/0.64	WITH/0.68	THEY/0.65	ON/0.72	THIS/0.89	BUT/0.84	IS/1.13	HE/0.88	IT/0.87
16	THIS/0.65	MONKEY/0.75	IS/0.59	THEY/0.68	SONG/0.65	THEY/0.71	NOT/0.86	I/0.8	HER/1	WITH/0.82	WITH/0.82
17	THEY/0.64	AS/0.72	SAID/0.58	SAID/0.68	THAT/0.63	HIM/0.69	IS/0.83	ON/0.77	NOT/0.97	HAD/0.74	THIS/0.79
18	PILGRIM/0.63	HIM/0.63	THAT/0.53	BUT/0.66	IT/0.6	FOR/0.66	THAT/0.8	WAS/0.75	THIS/0.84	ON/0.73	HAD/0.71
19	IT/0.6	ON/0.63	ON/0.53	TS'AO/0.64	FOR/0.56	THIS/0.58	SAID/0.79	THIS/0.74	Y/0.79	SAID/0.7	ON/0.7
20	IS/0.59	FOR/0.61	HAD/0.52	NOT/0.58	SAID/0.53	SONG/0.57	THERE/0.67	FOR/0.74	FOR/0.71	BE/0.65	AS/0.69
21	WE/0.55	THEY/0.59	HAVE/0.47	FOR/0.58	IS/0.46	THAT/0.57	HIM/0.67	BE/0.68	HAD/0.7	HIS/0.64	HIS/0.65
22	HAVE/0.52	BE/0.58	AT/0.46	AS/0.58	ME/0.45	IS/0.55	BUT/0.63	HAD/0.65	THEY/0.69	AS/0.6	BUT/0.62
23	ON/0.52	IS/0.57	BUT/0.45	HAD/0.56	BE/0.45	ALL/0.51	FOR/0.62	HIS/0.63	BE/0.66	HAVE/0.58	THEY/0.53
24	BE/0.51	ALL/0.56	YOUR/0.44	AT/0.54	HAD/0.43	HAD/0.5	HAVE/0.58	IS/0.61	ON/0.65	THIS/0.56	SO/0.53

中国学者应用语言学英语论文中的词块研究

塔里木大学 郑红红

提要：本文调查中国学者应用语言学英语论文中最常用的词块，并与国外学者进行对比，探讨中国学者应用语言学英语论文中词块的使用特点。本研究借助语料库检索软件提取出中国学术英语语料库（CAWEC）和国外学术英语语料库（FAWEC）中最常用词块，利用对数似然率检验二者使用词块的差异，并从功能和结构两方面分析了目标词块，最后归纳总结出中国学者在英语学术论文中词块的使用特点。研究表明：（1）中国学者英语学术论文中的词块使用情况特点主要是由词块使用不灵活造成的过度使用与使用不足；（2）对词块的功能性分析看出中国学者使用的功能形式相对简单，因为主题的限定导致了实词词块偏多；（3）对词块的结构分析看出中国学者使用的词块结构类型和外国学者相比只有一类不同，对介词结构的使用数量与外国学者无明显差异。

关键词：应用语言学英语论文、词块、语料库、中国学者

一、引言

中国的英语学习者在写作过程中总是感到困难重重：不知如何表达内容，或者先想中文后译成英语，或者文章条理混乱，内容空洞，搭配错误，词不达意等。英语学习者在写作过程中出现上述问题，是由于学习者大脑记忆中储存的相关词汇组合，短语结构等知识不够，在需要使用的场合捉襟见肘。语言学家把这种预置在脑海中的固定或半固定结构称为“词块”（chunks）。这正是Pawley & Syder（1983）指出的困扰二语习得者的两大问题，即如何获得接近本族语的流利性和接近本族语的选词能力。

在过去20年里，大型语料库和各种检索软件广泛应用于二语习得研究。基于语料库的研究表明，词块作为心理词库的基本单元，通常以整体存储或提取，从而无形中减轻了语言处理和输出的负担，使语言交际更加快捷、流利、有效（马广惠 2009）。从这个意义上说，词块与英语学习者的语言输出水平有着紧密的关系。Wray（1992）指出：“充分掌握一门新语言，需要学习者熟悉语言的本族语者更喜欢用哪些词组合。”

国内外研究者已经认识到词块在二语习得中的重要性。目前国内的词块研究也渐渐从最初的以理论探讨为主，转移到以英语学习者为对象的实证研究，但是对中国学者英语期刊论文中的词块研究涉及不多，所以需要相应的实证研究加以探讨。

二、文献综述

词块是以认知心理语言学为基础的。美国心理学家Miller（1956）提出了组块理论，人们可以借助于自己已有的知识和经历对信息进行组块和储存，扩大信息的容量，便于日后整体检索和提取。Sinclair（1991）认为语言习得包括两大体系：一个是语法规则为基础的分析性体系；另一个是以记忆为基础的套语体系。前者在记忆中所占空间小但强度大，即交际时难以准确、地道，后者包括大量的语块，即交际时易从记忆中提取，便于准确、流利表达。语言学家研究发现“语言并不是由语法和词汇组成，而是由大量的预制语块组成”。词块是比词语搭配更大的语言使用单位。语料库研究和分析表明，在自然语言中存在着大量的出现频率高、不同程度词汇化的词串，构成了英语中基本的单位。

Biber *et al.*（1999）对多字词语按习语、搭配、词汇语法关系和词块四个类别进行了简要讨论。他指出，词块为扩展搭配（extended collocation），是在语料中出现的词组合，可以是两词、三词、四词或四词以上的组合。Biber等人根据语料分析结果，把学术书面语中的词块从结构上分为12类（表1）。

表1.　学术书面语中的词块类型（Biber *et al.* 1999）

1. 名词短语+ of 短语片段结构	7. 系动词be + 名词短语/形容词片段
2. 名词短语+其他后修饰语片段	8.（动词短语+）that 从句片段
3. 介词短语+嵌带of 短语片段	9.（动词/形容词+）to 从句片段
4. 其他介词短语片段	10. 副词从句片段
5. 先行词 it +动词短语/形容词片段	11. 代词/名词短语+ be (+...) 片段
6. 被动动词+介词短语片段	12. 其他形式

国内对英语学习者的研究主要包含三个方面。1）关于词块对英语写作的重要性的非实证性探讨，比如词块在英语写作教学中的优势及可行的训练方法。2）以语料库为基础的实证研究。文秋芳、丁言仁、王文宇（2003）对比了英语专业四个年级英语学习者语料库，并与不同母语背景的英语学习者语料库作了比较，发现中国高水平英语学习者的书面语中表现出较强的口语化倾向，但它不受母语/文化背景差异影响，而是随着英语水平的提高，书面语中口语化倾向有弱化的趋势。王立非、张岩（2006）的研究显示中国学生在写作中存在以下情况：过度使用3词词块，词块的种类较少；过度使用名词语块和动词词块；中国学生使用的词块与本族语者有较大差异，具有口语化倾向；中国学生使用被动句式建构语块比本族语者少，使用主动句式多于本族语者。3）有关英语词块的使用与英语写作水平和英语作文质量的关系的实证研究。马广惠（2009）研究英语专业学生二语限时写作中的词块，卡方检验显示二语限时写作中目标词块的分布与英语说明文中的分布存在

显著差异；目标词块在英语说明文中出现的频率越高，在二语写作中的输出率也越大；在二语限时写作中，目标词块的特点：在二语限时写作中出现频率高，文本分布广泛。

此外，部分学者提出了检验同一词块在不同语料库中的频数是否具有显著性差异的方法。常见的有卡方检验和对数似然比。Dunning（1993）指出对数似然检验比卡方检验更可靠。

三、研究方法

本研究旨在调查中国学者公开发表的英语学术论文中词块的使用情况，并与国外学者公开发表的学术论文进行对比，寻找差异，并探讨中国学者的英语期刊论文中是否具有口语化倾向，故需回答下列研究问题。

1）中国学者英语学术论文中最常见的词块有哪些?

2）中国学者与外国学者对词块的使用是否存在差异?

3）如果差异存在，造成这些差异的原因有哪些?

为达成研究目标，笔者收集了《中国应用语言学》期刊中自2005年1月至2010年6月间的由中国学者撰写的英文论文，整理后建成中国学者学术英语语料库（Chinese Academic Written English Corpus，以下简称CAWEC）。作为对比参照的语料库为国外学者学术英语语料库（Foreign Academic Written English Corpus，以下简称FAWEC）。该语料库由马晓雷博士建成，语料来源于较为权威的国际性外语类期刊：*TESOL Quarterly*、*Language Learning*、*Applied Linguistics*和*SSLA*。《中国应用语言学》中发表的论文大都与英语教学相关，绝大多数是由中国英语教学方面的学者和教师撰写的（期刊中由外国学者撰写的文章被删去），而FAWEC中四本期刊是比较权威的国际期刊，主要内容也与英语教学相关，因此FAWEC可作为CAWEC的参照语料库。

语料库检索工具为WordSmith 5.0。对数似然率（Log-likelihood Ratio，缩写为LL）用来检验同一词块在不同语料库中的频数是否具有显著性差异。

基于语料库的实证研究中，词块被定义为每百万词（mw）中出现F次，同时有一定分布的N个词的组合，提取词块的频次，称为提取频点（cut-off frequency）；出现频次等于或大于提取频点的词组合，在文本分布数给定时，都被视为目标词块（马广惠2009）。Biber *et al*.（1999）按照频次为10次/mw、且至少分布在5个不同文本的标准提取目标词块，Cortes（2004）用的是“最保守的”20次/mw的提取点。可见，频次越高、文本分布越广的词块，越有研究价值和普遍意义。本研究采用马广惠（2009）的提取方法，按形符（tokens）提取，在中国学者学术英语语料库中目标词块的提取标准是，标准频数为40次/mw，且至少在5个文本中出现。首先使用WordSmith 5.0软件中的wordlist功能，每次改变设置中的cluster长度选项，分别提取中国学者学术英语语料库中3至6词目标词块词表，然后人工删除词表中的非词块语序列（不符合语法或者结构残缺的词组），从而获得最终数据。

四、结果

4.1 CAWEC中最常用的词块

为进一步探究中国学者学术英语词块的特点，由于每组词块的频数的不同，按比例将频率排名前20的3词词块、排名前10的4词词块、排名前5的5词块视为目标词块。中国学者学术英语中最常见的词块如表2所示。

表2. 中国学者学术英语中最常见的词块分布

分类	词块
3词词块	in order to, the use of, in terms of, as well as, the present study, one of the, teaching and learning, the process of, based on the, according to the, in this study, of the students, the target language, of the other, the results of, the meaning of, the other hand, the development of, in other words, use of the
4词词块	on the other hand, in the process of, at the same time, on the basis of, the results of the, at the end of, in the present study, the end of the, the meaning of the, in the use of
5词词块	at the end of the, at the beginning of the, and at the same time

4.2 对数似然率结果

4.2.1 CAWEC中最常见3词词块使用情况及对数似然率结果

表3. CAWEC中最常见3词词块及对数似然率结果

3词词块	CAWEC频数	CAWEC标准频率	FAWEC频数	FAWEC标准频率	LL	显著性
in order to	815	0.0504%	1,983	0.0337%	87.94	0.000
the use of	752	0.0465%	2,633	0.0447%	0.85	0.358
in terms of	630	0.0389%	2,223	0.0378%	0.45	0.502
as well as	521	0.0322%	2,364	0.0402%	-21.85	0.000
the present study	515	0.0318%	1,782	0.0303%	0.99	0.320
one of the	498	0.0308%	1,459	0.0248%	16.76	0.000
teaching and learning	464	0.0287%	232	0.0039%	650.39***	0.000

（待续）

（续表）

3词词块	CAWEC频数	CAWEC标准频率	FAWEC频数	FAWEC标准频率	LL	显著性
the process of	452	0.0279%	671	0.0337%	199.02	0.000
based on the	437	0.0270%	952	0.0162%	73.33	0.000
according to the	433	0.0268%	651	0.0111%	186.23	0.000
in this study	398	0.0246%	1,739	0.0295%	-11.30	0.001
of the students	384	0.0237%	362	0.0061%	320.57	0.000
the target language	379	0.0234%	833	0.0141%	61.64	0.000
of the other	331	0.0205%	289	0.0049%	299.36	0.000
the results of	331	0.0205%	1,430	0.0243%	-8.24	0.004
the meaning of	301	0.0186%	715	0.0121%	36.12	0.000
the other hand	297	0.0183%	297	0.0050%	232.12	0.000
the development of	290	0.0179%	960	0.0163%	1.95	0.163
in other words	285	0.0176%	1,194	0.0203%	-4.72	0.030
use of the	281	0.0174%	960	0.0163%	0.85	0.358

从表4.2中可看出，中国学者过度使用80%的最常用的3词词块，其中11个词块的使用频率与外国EFL学者有显著差异：in order to (LL=87.94, p<.01) , one of the (LL=16.76, p<.01), teaching and learning (LL=650.39, p<.01), the process of (LL=199.02, p<.01), based on the (LL=73.33, p<.01), according to the (LL=186.23, p<.01), of the students (LL=320.57, p<.001), the target language (LL=61.64, p<.001), of the other (LL=299.36, p<.01), the meaning of (LL=36.12, p<.01), the other hand (LL=232.12, p<.01)；有5个词块的使用频率高于外国EFL学者，但没有显著性差异：the use of (LL=0.85, p>.05), in terms of (LL=0.45, p>.05), the present study (LL=0.99, p>.05), the development of (LL=1.94, p>.05), use of the (LL=0.85, p>.05)；剩下的4个词块的使用频率低于外国EFL学者：as well as (LL=-21.85, p<.01), in this study (LL=-11.3, p=.01), the results of (LL=-8.24, p<.01), in other words (LL=-4.72, p<.01)。

4.2.2 CAWEC中最常用4词词块使用情况及对数似然率结果

最常见的4词词块可分为三类（如表4所示）：过度使用且差异显著、使用不足且差异显著及使用不足但差异不显著。

1）过度使用且差异显著：in the process of (LL=157.87, p<.01), at the same time

(LL=151.62, p<.01), the results of the (LL=128.95, p<.01), at the end of (LL=15.95, p<.01), the end of the (LL=11.18, p<.01), the meaning of the (LL=27.15, p<.01), in the use of (LL=25.66, p<.01)。

2）使用不足且差异显著：on the basis of (LL=-24.2, p<.01)。

3）使用不足但差异不显著：on the other hand (LL=-2.59, p>.05), in the present study (LL=-1.06, p>.05)。

表 4. CAWEC 中最常见的 4 词词块及对数似然率结果

4 词词块	CAWEC 频数	CAWEC 标准频率	FAWEC 频数	FAWEC 标准频率	LL	显著性
on the other hand	296	0.0183%	1,194	0.0203%	-2.59	0.108
in the process of	202	0.0125%	202	0.0034%	157.87	0.000
at the same time	194	0.0120%	194	0.0033%	151.62	0.000
on the basis of	172	0.0106%	925	0.0157%	-24.20	0.000
the results of the	165	0.0102%	165	0.0028%	128.95	0.000
at the end of	157	0.0097%	387	0.0066%	15.95	0.000
in the present study	155	0.0096%	618	0.0105%	-1.06	0.303
the end of the	143	0.0088%	371	0.0063%	11.18	0.001
the meaning of the	133	0.0082%	273	0.0046%	27.15	0.000
in the use of	122	0.0075%	248	0.0042%	25.66	0.000

4.2.3 CAWEC 中最常见的 5 词词块使用情况及对数似然率结果

表 5. CAWEC 中最常见的 5 词词块使用情况及对数似然率结果

5 词词块	CAWEC 频数	CAWEC 标准频率	FAWEC 频数	FAWEC 标准频率	LL	显著性
at the end of the	106	0.0065%	235	0.0040%	16.71	0.000
at the beginning of the	61	0.0038%	108	0.0018%	18.59	0.000
English as a foreign language	53	0.0033%	137	0.0023%	4.23	0.040
on the other hand the	51	0.0032%	176	0.0030%	0.11	0.742
and at the same time	47	0.0029%	46	0.0008%	37.64	0.000

如表5所示，CAWEC中最常见的5个5词词块均为过度使用，其中有3个与外国EFL学者有显著差异：at the end of the (LL=16.71, p<.01), at the beginning of the (LL=18.59, p<.01)，and at the same time (LL=37.64, p<.01)；对词块English as a foreign language (LL=4.23, p<.05)使用频率高于外国EFL学者，但显著性没有前三个词块高；对词块in the present study的使用频率高于外国EFL学者，但不存在显著差异。

从上述3个表中可清楚地看出，中国学者对4词词块和5词词块的使用少于3词词块，因此可推断出词块中所组合的单词越多该词块出现的频率就会越低。同时还能发现部分4词词块或5词词块中包含了3词词块，如the results of – the results of the, the other hand – on the other hand – on the other hand the，有个4词词块扩展成为了5词词块，如at the end of 和 at the end of the, at the same time 和 and at the same time。

4.3　中国学者词块使用的特点

4.3.1　功能层面

根据Biber *et al.*（1999）提出的基于功能层面的划分方法，可将词块分为四类：指示词块（referential chunks）、语篇组织词块（text organizers）、立场词块（stance chunks）和人际互动词块（interactional chunks）。

表6.　从功能角度划分CAWEC和FAWEC中的目标词块

类型	CAWEC	FAWEC
指示词块	the use of, the process of , in this study, in terms of, of the other, the results of, the meaning of, the development of, in other words, use of the, in the process of, the results of the, at the end of, the end of the, the meaning of the, in the present study, the meaning of the, at the end of the, at the beginning of the, in the use of the, of the students	the use of, in terms of, in this study, a number of, one of the, on the other, the number of, the role of, in which the, the acquisition of, part of the, the case of the, in other words, the effects of, in the case of, the result of the, in the present study, in the United States, the extent to which, the nature of the, in the context of, at the end of the, in the case of the, the results of
语篇组织词块	in order to, as well as, based on the, according to the, on the basis of, on the other hand, at the same time, on the other hand the, and at the same time	as well as, in order to, the fact that, on the other hand, on the basis of, at the same time, on the basis of the, on the other hand the
立场词块	/	it should be noted that
其他实词词块	the present study, teaching and learning, the target language, the other hand, the present study, English as a foreign language	the present study, the other hand

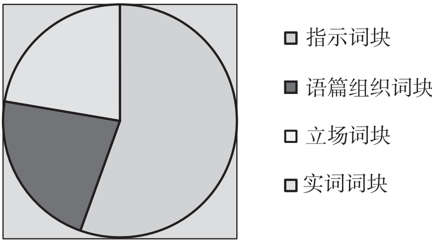


图 1. CAWEC 中各组功能性词块分布情况

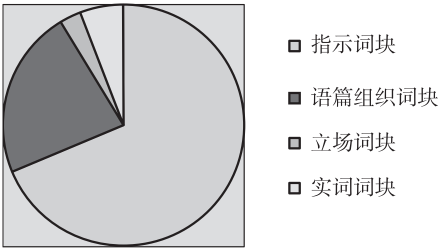


图2. FAWEC 中各组功能性词块分布情况

CAWEC的目标词块中包括20个指示词块和9个语篇组织词块，另外6个无功能性作用的实词词块。FAWEC的目标词块中有24个指示词块，8个语篇组织词块，1个立场词块及2个实词词块（见表6）。

综上所述，与国际学者相比，中国学者使用词块的功能形式较为简单且实词词块占的比例较大（图1和图2）。

4.3.2　结构层面

本研究根据Biber *et al.*（1999）对词块在结构层面的划分方法，将目标词块分为七类：1）名词短语+of 短语片段结构；2）名词短语+其他后修饰语片段；3）介词短语+嵌带of短语片段；4）其他介词短语片段；5）被动动词+介词短语片段；6）(动词短语+）that 从句片段；及7）其他形式。

表7. 从结构层面分析CAWEC与EAWEC中的目标词块

结构	CAWEC	FAWEC
1. 名词短语+of 短语片段结构	the use of, one of the, the process of, the results of, the meaning of, the development of, use of the, the results of the, the end of the, the meaning of the	the use of, a number of, one of the, the results of, the number of, the role of, the acquisition of, part of the, the case of, the effects of, the result of the, the nature of the
2. 名词短语+其他后修饰语片段	the present study, teaching and learning, the target language, the other hand, in other words, English as a foreign language	the present study, the fact that, the other hand, the extent to which
3. 介词短语+嵌带of短语片段	in terms of, on the basis of, at the end of, in the process of, in the use of the, at the end of the, at the beginning of the	in terms of, , in the case of, on the basis of, on the basis of the, in the context of, in the case of the, at the end of the

（待续）

（续表）

结构	CAWEC	FAWEC
4. 其他介词短语片段	in order to, according to the, in this study, of the other, of the students, on the other hand, at the same time, in the present study, on the other hand the, and at the same time	in order to, in this study, in which the, on the other, on the other hand, in the present study, at the same time, in the United States, on the other hand the, in other words
5. 被动动词+介词短语片段	based on the	/
6.（动词短语+）that 从句片段	/	It should be noted that
7. 其他形式	as well as	as well as

表 8. CAWEC 与 FAWEC 中目标词块结构分析比较

结构	CAWEC	FAWEC
1. 名词短语+ of 短语片段结构	10	12
2. 名词短语+其他后修饰语片段	6	4
3. 介词短语+嵌带 of 短语片段	7	7
4. 其他介词短语片段	10	10
5. 被动动词+介词短语片段	1	0
6.（动词短语+）that 从句片段	0	1
7. 其他形式	1	1

从表7和8可推断出：

（1）CAWEC和FAWEC中词块使用的结构数量一样多，但CAWEC中有“被动动词+介词短语片段（如based on the）”这一结构，FAWEC中没有；同样FAWEC中有“(动词短语+）that 从句片段(如it should be noted that)”这一结构而CAWEC中没有。

（2）中国学者对“介词短语+嵌带of 短语片段”和“其他介词短语片段”两种结构的使用与国外学者相同。

（3）中国学者倾向使用“名词短语+其他后修饰语片段”结构，这可能是由两个语料库中期刊的内容造成的，因为CAWEC中的文章主要与EFL教学相关，而FAWEC中包含四个期刊，内容相对要宽泛一些。

4.4 造成中国学者与国际学者词块使用差异的原因

中国学者英语学术论文中词块的使用特点较为复杂，使用过度和使用不足共存，主要是由用词不灵活引起的。在研究过程中，笔者发现，从两个语料库中提取的目标词块中有40%相同，而从CAWEC中随意挑出的词块在FAWEC中也存在。由此可见，中国学者具有较强的词块使用意识。此外，笔者检查了CAWEC中过度使用词块的具体使用情况，并与国际学者相比较，发现这些词块使用正确。部分学者提到造成中国学者使用词块较少的原因是由于中国学者有限的词汇量。而本研究中提到的中国学者大部分为研究英语教学方面的专家和大学外语教师，他们已具较高的英语水平，因此有限的词汇量不能完全解释中国学者过多或过少使用词块。

在中国学者学术英语论文中，之前使用的词块会在后文中反复使用，因此中国学者使用词块的类型更为单一，而国际学者在表达相同意思时更倾向于使用不同形式的词块。例如，中国学者对based on the过度使用，但对表示相同意思的on the basis of使用不足。有一些词块的过度使用可能是受期刊话题的影响，如teaching and learning和the target language等。

5. 结论

本研究借助语料库技术研究了中国学者学术英语论文中3词、4词、5词块词块的使用情况，并与国际学者进行比较，发现中国学者的英语学术论文中的词块具有过度使用和使用不足的特点，主要是由用词不灵活引起的。但总体而言，研究结果表明中国学者对词块的使用把握较好，可为英语专业研究生和EFL教师提供有针对性的帮助。

研究者应注重培养词块意识，在学习使用词块的过程中不应只注重增加词汇量，应关注词块在目标语言中的用法，灵活使用词块，不断提高英语水平，争取达到本族者的水平。而英语写作教学过程中，EFL教师则应多向学生教授不同类型的连接词和话语标记语，鼓励学生使用词块，培养学生在写作时注重文章的连贯性和逻辑性的意识。

本研究所使用的语料库CAWEC和FAWEC中的语料话题都与英语教学相关，但CAWEC中的语料来源于一本期刊，而FAWEC中的语料来源于4本期刊，因此FAWEC中的语料话题比CAWEC要更为宽泛。另外，CAWEC有100万字而FAWEC中有350万字，但在研究中只将频率最高的35个词块作为目标词块，因此具有一定的局限性。今后的研究可拓展到英语专业研究生撰写的英语学术论文的研究。

参考文献

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Spoken and Written English* [M]. New York: Longman.

Cortes, V. 2004. Lexical bundles in published and student disciplinary writing: Examples from

history and biology [J]. *English for Specific Purposes* 23(4): 397-423.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence [J]. *Computational Linguistics* 19(1): 61-74.

Miller, G. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information [J]. *The Psychological Review* 63(2): 81-97.

Pawley, A. & F. Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency [A]. In J. Richards & R. Schmidt (eds.). *Language and Communication* [C]. London: Longman. 191-226.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.

Wray, A. 1992. *The Focusing Hypothesis: The Theory of Left Hemisphere Lateralized Language Re-examined* [M]. Amsterdam: John Benjamins.

马广惠，2009，英语专业学生二语限时写作中的词块研究 [J],《外语教学与研究》(1)：54-60。

王立非、张　岩，2006，基于语料库的大学生英语议论文中的语块使用模式研究 [J],《外语电化教学》(4)：36-41。

文秋芳、丁言仁、王文宇，2003，中国大学生英语书面语中的口语化倾向——高水平英语学习者语料库对比分析 [J],《外语教学与研究》(4)：268-274。

通信地址：843300 新疆阿拉尔市塔里木大学人文学院

语义韵研究的理论、方法与应用*

扬州大学 陆 军

提要：本文回顾和梳理了语义韵研究的理论演变、研究方法和应用研究。在此基础上剖析了各流派理论的形成机制、语言学贡献以及相互之间的关系；探讨了各研究方法的理论基础、优缺点及其改良途径和发展方向；讨论了各类应用研究的工作机制、理论贡献和发展前景。最后总结全文：语义韵研究的理论和方法体系基本确立，应用研究趋于朝纵深方向发展，对语言学和应用语言学研究乃至于认知科学研究具有重大潜在意义。

关键词：语义韵、理论演变、研究方法、应用领域

语义韵（semantic prosody）研究起源于20世纪80年代。Sinclair（1987，1991）基于语料库数据发现，动词短语SET in倾向于与表示消极事件的主语共现，形成带有消极语义氛围的语境，这标志着语义韵研究的萌芽。Semantic prosody这一术语由Louw（1993）首次正式使用。其中，prosody一词源于Firth的音位学研究，而semantic prosody则反映了语言交际中的词语组合所构成的态度意义，具有超越一定语言单位的特点。这种态度意义主要表达说话者的态度（Louw 2000：58），体现了短语单位与特定功能的共选，在词汇语法组合中起决定作用（Sinclair 1996，2004；Stubbs 2009）。

近20年来，基于语料库证据的语义韵研究日益兴起。Stewart（2010）出版了第一部语义韵研究专著《语义韵——批判性评估》，这标志着语义韵研究已步入新阶段。同时，其理论与研究方法等方面也日趋复杂化和体系化。因此，现阶段探讨语义韵理论、研究方法和应用研究，梳理和确立相关研究体系具有重要意义。

1. 语义韵理论

语义韵概念的界定和阐释，主要经历由“语义传染说”、“内涵意义说”到“功能说”的演变和深化（卫乃兴 2011a），反映了研究者对语义韵的形成机制、承载单位和实现方式的认识。

1.1 语义传染说

“语义传染说”多采用语义“转移”、“感染”和“附着”等表述方式，旨在强调语境

* 本研究系江苏省社科基金项目“语料库驱动的隐性、显性知识接口研究”（13YYB006）和扬州大学2012年度“新世纪人才工程”项目的阶段性成果。

中的意义流动，体现搭配群共享的语义跨越词语界限、感染节点词的特征。Louw（1993）指出，像SET in和symptomatic of等节点词的出现向读者（或听众）预示着带有特定语义特征的搭配词出现，这就是语义韵在起作用。究其原因，这些节点词习惯性地吸引某一类具有相同或相似语义特点的搭配词，它们在文本中反复共现，使得节点词被传染上相关语义特点，并使其整个跨距内弥漫着一种特殊的语义氛围即语义韵（同上：157）。Bublitz（1996）认为，这种节点词与搭配词习惯性共现形成特定语义氛围的原因在于："在同一语境中反复使用某个词最终会产生语义转移，即该词从邻近词语获得某些语义特征"（同上：11）。随着使用频率的增加，感染或转移的语义特征会越来越明显（Stubbs 1995：50）。由此说明，语义韵意味着搭配词与节点词间的意义转移，这种转移是经过一段时间的语义"感染"才产生的。正是由于这种感染效应，一旦带有不同语义特征的异质词出现，就会与人们所期待的语义氛围发生冲突，即语义韵冲突（prosodic clash），常常表现为反语（irony）等语言现象（Louw 1993：164）。

"语义转移"、"语义感染"和"语义附着"等特点反映了语义韵的形成过程，折射出其形成机制和物质基础，也反映了一些缺陷。其中，共选（co-selection）是主要形成机制，即人们在语言交际中倾向于选择某一类或几类语义特征的词语与某节点词搭配，它们反复共现形成特定的语义韵。这类词语组合总是和某一（些）语义韵共选。其中弥漫的语义氛围使得节点词看似失去部分意思又获得新的意思。人们之所以在交际中如此共选,是为了表达他们对特定事物或活动持有特别的态度，而表达这些态度又离不开特定的词语组合。因此，人类的主观评价与客观事件交互作用，构成了语义韵形成的物质基础。不过，"传染说"未能厘清语义韵和语义选择趋向（即搭配词语义特征）之间的本质区别。

1.2 内涵意义说

把语义韵视为内涵意义的观点在相关研究中较为普遍（Whitsitt 2005；Stewart 2010）。例如，Berber-Sardinha（2000：93）将语义韵定义为"习惯性共现的词语组合所表达的内涵意义"。这类观点认为，内涵意义是语言评价手段之一，语义韵也用于评价事物的好坏，二者都可能表现为潜在的、不为直觉所感知的评价意义。这些研究趋于把语义韵等同于或部分地等同于词语的内涵意义，侧重于强调语义韵是词语内涵意义在搭配型式中的延伸（如Hunston 2002；Partington 1998，2004；Stubbs 2001a等）。

其中，Partington（1998：67）认为，像COMMIT等词的消极内涵意义不仅存在于这些词本身，还在于其与搭配词所构成的短语单位之中。她还指出，语义韵是内涵意义的一个方面，不过往往要延伸到一个语言单位以外，不易为研究者直接觉察（Partington 2004：131-132）。类似地，在Stubbs（2001a）中有如是表述："内涵意义用于表达讲话者个人的态度，……，是核心意义中最重要的成分"（同上：35），"语义韵（discourse prosody）用于表达说话者的态度，……，是说话者选择某种表达的原因"（同上：65）和"需要依靠直觉来区别内在命题意义与内涵意义（或语义韵）"（同上：106）等。Hunston（2002：141-142）指出："语义韵解释了内涵意义，即词语所承载的'真实'意义以外的意义"。这些表述似乎都暗示着内涵意义和语义韵之间存在同义关系。

不过，学界对两者的关系持有不同见解。例如，Louw（2000：50）指出，语义韵并不只是内涵意义。内涵意义与重复事件的图式知识相关，而语义韵的本质是功能的、态度的。卫乃兴（2011a）强调，语义韵在很大程度上超越内涵意义，是语境层面上暗藏的态度意义。Sinclair（1996，2004）主要使用attitudinal描述语义韵，几乎从未使用connotational一词。他主张对共现语境中所蕴含的语用意义和态度意义进行具体描述，而不只是积极或消极。

根据共选模型（参见Sinclair 1996；Stubbs 2009）可以认为，语义韵和内涵意义之间存在本质联系，但属不同概念范畴。它们源于同一语言现象，但分属短语意义单位和单词意义单位。从短语意义单位看，词语组合共同实现某一态度意义即语义韵。然而，单词意义单位的观念根深蒂固，人们习惯性地赋予单个词特定的意义。根据词语结伴说（You shall know a word by the company it keeps, Firth 1957)，人们在交际中首先获取短语的意思，然后将之分配给或赋予单个词。然而，记忆中往往存储着同一节点词的多个词语组合（Hoey 2005），且它们所承载的态度意义具有隐藏特性，因此语义“分配”或“赋值”过程有很大的主观性和任意性。由此可见，根据语义韵所获得的单个词的言外之意一方面具有语义韵的某些特征，另一方面又会受到多种搭配词语义特征的影响而趋于模糊。再者，在分析研究意义时，内涵意义这一概念往往先入为主。因此，语义韵的界定和描述往往摆脱不了其影响。

1.3 功能说

以Sinclair 为代表的语义韵“功能说”揭示了语义韵与态度意义和交际意图紧密联系的本质特征，得到Tognini-Bonelli（2001）、Hunston（2002，2007）和Stubbs（2001a，2009）等赞同。他指出，语义韵更接近于“语义—语用”连续体的语用一侧（Sinclair 1996：87），主张将语义韵置于扩展意义单位模型中理解。该模型由核心词、词语搭配、类联接、语义选择趋向和语义韵五个要素共同界定。其中，语义韵是必需要素，表达了整个意义单位的交际目的和功能（Sinclair 1996；Stubbs 2009）。Tognini-Bonelli（2001）倾向于把语义韵的语用功能置于核心位置。她指出，“说话者或作者在选择一个多词单位时，不仅要考虑到一个词相邻位置的词语和语法共选关系和限制关系，还要考虑到更远的语义选择趋向和相应的语用关系即语义韵”（同上：111）。其中，语义韵是话语选择的依据，在共选和限制关系中起决定性作用，确立了具有一定功能的语篇单位（functional discourse units）。

Louw（2000：60）指出“语义韵的功能是用于表达说话者或作者对具体语用情景的态度，还用于创造讽刺效果”。Stubbs（2001a：66）认为“语义韵表达评价意义，是话语选择的原因和实现话语功能的单位”。他提出用discourse prosody替代semantic prosody来表示“超过一个语言单位、表达说话者态度的意义特征，……，这既能保持与说话者和听者之间的联系，也强调了其语篇衔接功能”（Stubbs 2001a：66）。由此可见，语义韵所实现的功能还包括意义单位在语篇构建中的衔接作用。这进一步肯定了Sinclair所强调的语用功能和语篇意义。

综上所述，语义韵“功能说”可视为从语义韵在交际中的具体实现以及其在语篇衔接

中的作用进行描述和界定。Sinclair（1996）和Tognini-Bonelli（2001）侧重于探讨短语单位内各要素如何交互作用实现具体的交际功能，从而确立语义韵在词汇语法组合中的统领作用，即语义韵是意义单位的核心成分，具有评价特征，是话语选择的首要依据，与内涵意义相区别。Louw（2000）和Stubbs（2001a）则进一步将语义韵的功能延伸至语言表达与语用情景之间的关系，凸显了其在语篇构建中的作用。

由此，“功能说”可视为在“语义传染说”和“内涵意义说”基础上对语义韵的进一步提炼，是更高层次的抽象和概括，蕴含着其在语言研究中的重要地位。“功能说”不但揭示了语义韵与词汇和语法的本质联系，还发现了其在语言交际中的核心作用：它不仅统领着词汇和语法项的组合，实现特定的意义和功能，构成短语单位，还负责着语篇中短语单位的相互衔接。这一方面从语言交际或语用角度论证了语义韵的形成机制，从共时角度阐释了语义韵“传染说”的理据，同时也厘清了其与内涵意义的本质区别。另一方面，“功能说”还奠定了语义韵至上的语言学意义：既然语义韵制约着交际中的词汇和语法选择，影响着语篇的衔接和连贯，所以无论是词汇、语法和语篇等语言现象的描述，还是相关语言现象的心理加工或认知机制的探索等都把语义韵放在首要位置。

2. 语义韵研究方法

语义韵研究主要采纳数据驱动的方法（data-driven approach）和基于数据的方法（data-based approach）。此外，基于扩展意义单位模型的多重比较法逐步被采纳。

2.1 数据驱动的语义韵研究方法

数据驱动的研究利用统计手段计算和提取搭配词，根据显著搭配词语义特征归纳语义韵。该方法以语义韵“感染说”为主要理论基础，与“内涵意义说”关系紧密。如1.1节所述，节点词容易受搭配词语义的感染而产生特定的语义氛围即语义韵。由此，可以通过观察高频搭配词语义特征（如积极或消极语义）归纳语义韵特征，如Tognini-Bonnelli（2001）、Xiao & McEnery（2006）和卫乃兴（2002c）等。其中，积极或消极语义特征与词语的内涵意义密切关联。具体操作中，往往借助于计算机软件（如WordSmith Tools）提取一定跨距内的显著搭配词用于考察。这些搭配词与节点词未必具有语法上的直接限制关系，但其语义特征在一定程度上代表了节点词语境内所弥漫的语义氛围。例如，Xiao & McEnery（2006）借助于频数和MI值等统计指标筛选出中英文近义词的显著搭配词，然后归纳语义韵特征。

此类研究主要依靠计算机程序进行自动统计测量、检索和提取，适用于大型语料库研究。此过程中人为因素的影响较少，有助于较为客观地揭示语义韵特征。然而，由于该方法更适用于搭配词语义特征较为明显的节点词，主要用于积极和消极语义韵研究，如Louw（1993，2000）、Stubbs（1995，1996，2001a，2001b）、Partington（1998）、Hunston（2002）和Schmitt & Carter（2004）等。然而，搭配词语义特征的典型性与节点词的搭配范围密切相关。搭配范围越大，所提取的显著搭配词的代表性就越小，语义韵归纳的信度和效度都会受到影响。

2.2 基于数据的语义韵研究方法

基于数据的语义韵研究方法以“功能说”为主要理论基础。它通过考察更丰富的索引行语境信息来确立短语意义单位的构成要素，然后根据各要素的特征概括短语单位的态度意义。如1.3节所述，“功能说”视语义韵为短语单位或扩展意义单位的交际目的和功能。实际语言运用中的交际目的往往更为微妙，远不止积极和消极两大类。基于数据的研究方法有助于考察更为具体的交际目的或功能。具体操作步骤包括：首先，利用统计抽样手段（如随机抽样）从语料库中提取足够数量的索引行；其次，观察索引行并确定类联接；再次，参照类联接检查和概括搭配词语义特征；最后，归纳语义韵。Sinclair（1996, 1998）是该方法的典型应用实例，分别以naked eye、true feelings、brook和my place等为节点词，归纳出“困难”、“不情愿”、“无能力”和“非正式邀请”等具体的态度意义。类似的例子参见卫乃兴（2002a, 2002b, 2011b）、李晓红、卫乃兴（2012a, 2012b）和陆军、卫乃兴（2012）等。

与数据驱动的方法相比，该方法有助于更为具体、精准地归纳出态度意义。这种态度意义可以隶属于“积极、消极”二分法以下。例如，“林奈双名法”（Linnaean-style binomial notation）就包括了宏观和微观两个层面的功能认定，如[积极：赞成]等。宏观层面为基本态度意义，分为“积极”和“消极”两类，微观层面可有多种具体的态度意义或功能描述，该层面所包含的信息比较全面（参见Morley & Partington 2009：141）。

2.3 多重比较法

上述方法都面临着这一难题：如何确保语义韵归纳的精准性和统一性？这直接制约着语义韵研究的深入开展，特别是妨碍了二语语义韵的型式特征、心理加工机制和认知规律等问题的解决。多重比较法以扩展意义单位模型为框架，从类联接、词语搭配、语义选择趋向和语义韵等方面描述语言型式，构成比较分析的多个要素；然后以这些要素为对象，进行二语（如中国学习者英语）、目标语（如英语本族语）和母语（如汉语本族语）之间的多重比较（参见陆军 2012；陆军、卫乃兴 2013）。其中，各要素层面的多重比较能为揭示二语语义韵形成机制提供具体证据；母语与目标语的异同可以预测二语语义韵的偏离方向；二语与目标语之间的异同能够反映二语语义韵的偏离程度；二语与母语的趋同性则为揭示母语影响提供证据。由此，尽管语义韵归纳缺乏统一标准，多重比较法能为精准廓清二语语义韵的相对特征提供丰富的证据。

严格意义上讲，多重比较法并非是与数据驱动或基于数据的方法的简单并列，而是基于数据的方法在对比短语学领域的具体发展。研究方法往往建立在一定的理论基础上，而理论发展又离不开具体的研究方法，二者相互促进，密不可分（参见Stubbs 2009）。数据驱动的方法和基于数据的方法都基于语言使用的概率属性，分别以“语义传染说”和“功能说”为主要理论基础，同时也促进了语义韵理论的发展。多重比较法以“共选”特别是“功能说”为理论基础的同时，还整合了“对比分析”（CA）和“中介语对比分析”（CIA）的理论成果（参见Granger 1998：14），对语义韵研究的专业化和精确化具有启示意义，反

映了语义韵研究方法的发展方向。

上述方法各有优缺点。其中，数据驱动的方法适合大规模数据的自动提取和处理，但不利于揭示具体的态度意义或语用功能；基于数据的方法有助于发现微妙的态度意义，但未必能反映具有统计意义的词语组合或短语单位，且具体态度意义认定的难度很大，不易在大规模研究中操作。多重比较法有助于克服态度意义认定困难的问题，但涉及多方面的数据，聚焦于对比研究，研究成本较高，适用范围相对较窄。不过，在实际研究中可将这些方法有机结合，如可使用数据驱动的方法来验证另外两种方法的研究发现，这样有助于得到更可靠的研究结果（参见 Tognini-Bonelli 2001：24）。

3. 语义韵研究领域

随着语义韵理论和研究方法的深入探讨，语义韵研究已从单语文本研究扩展到多语对比研究等领域，主要可概括为：1）单语环境下的语义韵特征描述和影响因素探讨；2）双语语义韵对比分析；3）二语语义韵特征探索。

3.1　单语文本中的语义韵特征研究

早期的语义韵研究在单语环境下开展，主要利用大型通用语料库调查语义韵特征和探讨研究方法，如Sinclair（1991，1996，1998）、Stubbs（1995，1996，2001a，2001b）和Stewart（2003）等。在此基础上逐步探讨出“数据驱动的”和“基于数据的”语义韵研究方法。其中，近义词语义韵的比较研究是一个亮点，如Partington（1998）等。研究者注意到，不同词语（如近义词）在语义韵上存在差异，即使同一词语在不同类型的文本中也会构筑不同的语义韵。于是，特定类型文本的语义韵特征备受关注，文学语篇尤为典型，如Louw（1993）、Adolphs & Carter（2002）、Stubbs（2005）、Adolphs（2006）和O’Halloran（2007）等。

相应地，专业文本的语义韵研究也逐步受到关注，倾向于以通用语料库为参照。例如，Tribble（2000）考察了欧盟项目建议书语料库中EXPERIENCE的语义韵特征。卫乃兴（2002a）利用JDEST语料库调查了学术文本中CAUSE、CAREER和PROBABILITY所形成的语义氛围。Nelson（2006）概括了COMPETITIVE、MARKET和EXPORT在商务语篇中所实现的语义韵特征，并由此注意到专业文本特有的搭配形成过程。Cheng（2006）对SARS文本进行考察，试图说明习惯性词语共选对语篇意义和连贯的累积效应。

此类研究揭示了直觉未能直接觉察到的语义韵特征。这些特征与文本的专业领域密切相关（Hunston 2007）。“在特定的文类中，某些词可能会构筑该类文本特有的语义韵，即局部语义韵（local prosody），但并非所有关键词都这样”（Tribble 2000：86）。这些研究为语义韵“传染说”、“内涵意义说”和“功能说”的发展奠定了基础，有助于确立语义韵理论体系。

3.2　跨语言语义韵特征对比研究

随着语义韵理论和研究方法趋于完善，跨语言语义韵对比研究日渐兴起。主要包括不

同语言中对应词语的语义韵对比研究。例如，Partington（1998）通过语义韵分析揭示了英语和意大利语中的“假朋友”（false friends）现象；Berber-Sardinha（2000）和Tognini-Bonelli（2001）分别开展了英语与葡萄牙语和英语与意大利语的语义韵对比研究；Xiao & McEnery（2006）比较分析了英汉近义词语义韵特征等。张继东、刘萍（2006）比较了英汉动词happen、occur和“发生”的语义韵。研究发现，即使是翻译对应词语在语义韵上也可能存在很大差异。

上述发现激发了研究者对跨语言语义韵对应机制等问题的兴趣。翻译活动和翻译文本的语义韵特征自然成为对比研究关注的对象。Dam-Jensen & Zethsen（2008）考察了非本族语英语学习者在翻译中的语义韵意识；Stewart（2009）调查了英意翻译中的语义韵问题。Wei & Li (2014)、卫乃兴（2011a，2011b）、李晓红、卫乃兴（2012a，2012b）和陆军、卫乃兴（2012）等揭示了英汉对应词的语义选择趋向和语义韵特征及其它们在实现跨语言对应中的作用，从而探讨了英汉对应机制等问题，建立了基于扩展意义单位模型的对比短语学研究框架（参见卫乃兴、陆军2014）。跨语言对比研究发现，语义韵是翻译中功能对等机制的核心内容，这从双语角度证实了语义韵在词汇和语法组合中的统领作用。

3.3　二语语义韵研究

二语语义韵研究也逐步成为热点，主要考察二语学习者语言与目标语在语义韵特征上的异同和变化规律。与3.2节中的跨语言对比研究相比，此类研究常使用目标语语料库作为参照语料库。近年来，我国学者参照英语本族语语料库针对中国学习者英语语义韵特征开展了大量研究（参见翟红华、方红秀 2009），如孙海燕（2004）、王海华、王同顺（2005）、卫乃兴（2006）、黄瑞红（2007）、王春艳（2009）、陆军（2010，2012）和陆军、卫乃兴（2013）等。研究发现，中国学习者英语与英语本族语在语义韵上可能存在明显差异（差异大小与英汉语本族语差异密切相关），但与汉语本族语趋于高度一致。由此说明：母语语义韵知识在二语知识体系构建中起主导作用。这从二语角度揭示了词汇语法共选机制中语义韵起核心作用，对二语心理加工和认知规律探索具有启示意义。

4. 结语

综上所述，语义韵研究在理论探讨、方法建设和应用研究等方面都取得长足的发展，同时也都面临着新的挑战和发展机遇。

首先，语义韵理论体系构建趋于系统化，但迫切需要建立语义韵分类标准。“语义传染说”、“内涵意义说”和“功能说”从不同角度对语义韵进行描述和界定。这反映了语义韵理论体系的复杂性，有其独特的物质基础（短语单位与特定交际目的共选）、概念体系（与“内涵意义”相区别）和形成机制（语义传染）。这种复杂性使得其在概念范畴上存在分歧。大量研究都指出，语义韵是潜意识的，不易为直觉所觉察，将隐性特征视为语义韵的根本特性，如Louw（1993：169-171）、Tognini-Bonelli（2001：112）、Hunston（2002：21）、Partington（2004：131）和McEnery *et al.*（2006：84）等。然而，根据

Schmitt *et al.*（2001）对意识的分类可知，隐性特征存在一个程度问题。正是由于不同程度的隐性特征，微观层面的态度意义难以统一，迫切需要建立满足语义韵研究需要的态度意义体系。这直接制约着像“是否所有词或短语都实现语义韵?”等重要问题的探讨（参见Stubbs 2009：29）。

其次，语义韵归纳方法主要基于语言的概率属性，但面临着认知研究的挑战。基于概率的语义韵研究方法的确立反映了语义韵具有语言的根本属性，语义韵既是一种普遍的语言现象，同时也是一种抽象程度高于词汇和语法的要素，在语言表达中起统领作用。然而，语义韵在多大程度上或在何种情况下能被人的直觉或潜意识所发现等问题，语料库数据并不能够回答。可以基于语料库数据形成相应的推论（如语义韵的“隐性”认知特征），但推论的准确性还需要借助于认知学、心理学的研究方法来验证。

第三，语义韵研究正朝纵深方向发展，但需要开拓更广阔的应用前景。单语语境下的语义韵描述揭示了语义韵的特征、形成机制和交际功能；双语语义韵特征的异同在一定程度上反映了双语交际中语义韵的作用方式；二语与目标语在语义韵特征上的偏差更微妙地反映了语义韵在词汇语法共选机制中的核心作用。由此可以认为，语义韵研究正从语言本体的描述向语言认知加工研究延伸。事实上，既然语义韵是在语言交际中起主导作用的语言现象，它必然会受到语言学其他领域的关注，如语义韵如何被人脑认知、加工和处理等问题仍需要解释。与此同时，语义韵的交际主导作用本身也说明，它比其他语言现象更加能够对心理处理和认知加工研究贡献力量。这反映了其在语言科学和认知科学等研究领域的广阔应用前景。

参考文献

Adolphs, S. 2006. *Introducing Electronic Text Analysis* [M]. London: Routledge.

Adolphs, S. & R. Carter. 2002. Points of view and semantic prosodies in Virginia Woolf's *To the Lighthouse* [J]. *Poetica* 58: 7-20.

Berber-Sardinha, T. 2000. Semantic prosodies in English and Portuguese: A contrastive study [J]. *Cuadernos de Filología Inglesa* 9(1): 93-110.

Bublitz, W. 1996. Semantic prosody and cohesive company: Somewhat predictable [J]. *Leuvense Bijdraden: Tijdschrift voor Germaanse Filologie* 85(1-2): 1-32.

Cheng, W. 2006. Describing the extended meanings of lexical cohesion in a corpus of SARS spoken discourse [J]. *International Journal of Corpus Linguistics* 12(3): 325-344.

Dam-Jensen, H. & K. Zethsen. 2008. Translator awareness of semantic prosodies [J]. *Target* 20(2): 203-231.

Firth, J. 1957. A synopsis of linguistic theory, 1930-55. Studies in Linguistic Analysis [A]. Reprinted in F. Palmer. 1968. *Selected Papers of J. R. Firth 1952–59* [C]. Bloomington: Indiana University Press. 168-205.

Granger, S. (ed.). 1998. *Learner English on Computer* [C]. London: Longman.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language* [M]. London: Routledge.

Hunston, S. 2002. *Corpora in Applied Linguistics* [M]. Cambridge: CUP.

Hunston, S. 2007. Semantic prosody revisited [J]. *International Journal of Corpus Linguistics* 12(2): 249-268.

Louw, B. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies [A]. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.). *Text and Technology: In Honour of John Sinclair* [C]. Amsterdam: John Benjamins. 157-176.

Louw, B. 2000. Contextual prosodic theory: Bringing semantic prosodies to life [A]. In C. Heffer, H. Sauntson & G. Fox (eds.). *Words in Context: A Tribute to John Sinclair on his Retirement* [C]. Birmingham: University of Birmingham.

McEnery, T., R. Xiao & Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book* [M]. London: Routledge.

Morley, J. & A. Partington. 2009. A few frequently asked questions about semantic - or *evaluative*-prosody [J]. *International Journal of Corpus Linguistics* 14(2): 139-158.

Nelson, M. 2006. Semantic associations in business English: A corpus-based analysis [J]. *English for Specific Purposes* 25: 217-234.

O'Halloran, K. 2007. Corpus-assisted literary evaluation [J]. *Corpora* 2(1): 33-68.

Partington, A. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching* [M]. Amsterdam: John Benjamins.

Partington, A. 2004. "Utterly content in each other company" : Semantic prosody and semantic preference [J]. *International Journal of Corpus Linguistics* 9(1): 131-156.

Schmitt, N., D. Schmitt & C. Clapham. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test [J]. *Language Testing* 18(1): 55-88.

Schmitt, N. & R. Carter. 2004. Formulaic sequences in action: An introduction [A]. In N. Schmitt (ed.). *Formulaic Sequences* [C]. Amsterdam: John Benjamins. 1-22.

Sinclair, J. 1987. *Looking Up* [C]. London: Collins.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.

Sinclair, J. 1996. The search for units of meaning [J]. *Textus* (4): 75-106.

Sinclair, J. 1998. The lexical item [A]. In E. Weigand (ed.). *Contrastive Lexical Semantics* [C]. Amsterdam: John Banjamins. 124.

Sinclair, J. 2004. *Trust The Text* [M]. London: Routledge.

Stewart, D. 2003. The corpus of revelation: The BNC, semantic prosodies and syntactic shells [A]. In B. Lewandowska-Tomaszczyk (ed.). Proceedings of the conference *Practical Applications in Language Corpora* (PALC), Lods, 7-9 September 2001 [C]. Frankfurt: Peter Lang. 329-341.

Stewart, D. 2009. Safeguarding the lexicogrammatical environment: Translating semantic prosody [A]. In A. Beeby, P. Rodriguez-Ines & P. Sanchez-Gijon (eds.). *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate* [C]. Amsterdam: John Benjamins. 29-46.

Stewart, D. 2010. *Semantic Prosody: A Critical Evaluation* [M]. London: Routledge.

Stubbs, M. 1995. Collocations and semantic profiles：On the cause of the trouble with quantitative studies [J]. *Functions of Language* 2(1): 23-55.

Stubbs, M. 1996. *Text and Corpus Analysis* [M]. Oxford: Blackwell.

Stubbs, M. 2001a. *Words and Phrases: Corpus Studies of Lexical Semantics* [M]. Oxford: Blackwell.

Stubbs, M. 2001b. On inference theories and code theories: Corpus evidence for semantic schemas [J]. *Text* 21(3): 437-465.

Stubbs, M. 2005. Conrad in the computer: Examples of quantitative stylistic methods [J]. *Language and Literature* 14(1): 5-24.

Stubbs, M. 2009. John Sinclair (1933–2007): The search for units of meaning: Sinclair on empirical semantics [J]. *Applied Linguistics* 30(1): 115-137.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [M]. Amsterdam: John Benjamins.

Tribble, C. 2000. Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals [A]. In L. Burnard & A. McEnery (eds.). *Rethinking Language Pedagogy from a Corpus Perspective* [C]. Frankfurth: Peter Lang. 75-90.

Wei, N. & X. Li. 2014. Exploring semantic preference and semantic prosody across English and Chinese: Their roles for cross-linguistic equivalence [J]. *Corpus Linguistics and Linguistic Theory* 10(1): 103-138.

Whitsitt, S. 2005. A critique of the concept of semantic prosody [J]. *International Journal of Corpus Linguistics* 10(3): 283-305.

Xiao, R. & T. McEnery. 2006. Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective [J]. *Applied Linguistics* 27(1): 103-129.

黄瑞红，2007，中国英语学习者形容词增强语的语义韵研究 [J],《外语教学》(4)：57-60。

李晓红、卫乃兴，2012a，汉英对应词语单位的语义趋向及语义韵对比研究 [J],《外语教学与研究》(1)：20-33。

李晓红、卫乃兴，2012b，双语视角下词语内涵义与语义韵探究 [J],《现代外语》(1)：30-38。

陆 军，2010，基于语料库的学习者英语近义词搭配行为与语义韵研究 [J],《现代外语》(3)：276-286。

陆 军，2012，共选理论视角下的学习者英语型式构成特征研究 [J],《现代外语》(1)：70-78。

陆 军、卫乃兴，2012，扩展意义单位模型下的英汉翻译对等型式构成研究 [J],《外语教学与研究》(3)：424-436。

陆 军、卫乃兴，2013，共选视阈下的二语知识研究：一项语料库驱动的使役态共选特征多重比较 [J],《外语界》(3)：2-11。

孙海燕，2004，基于语料库的学生英语形容词搭配语义特征探究 [J],《现代外语》(4)：410-419。

王春艳，2009，基于语料库的中国学习者英语近义词区分探讨 [J],《外语与外语教学》(5)：27-31。

王海华、王同顺，2005，CAUSE语义韵的对比研究 [J],《现代外语》(3)：297-307。

卫乃兴，2002a,《词语搭配的界定与研究体系》[M]。上海：上海交通大学出版社。

卫乃兴，2002b，语义韵研究的一般方法 [J],《外语教学与研究》(4)：300-307。

卫乃兴，2002c，语料库驱动的专业文本语义韵研究 [J],《现代外语》(2)：165-175。

卫乃兴，2006，基于语料库学生英语中的语义韵对比研究 [J],《外语学刊》(5)：50-55。

卫乃兴，2011a，《词语学要义》[M]。上海：上海外语教育出版社。
卫乃兴，2011b，基于语料库的对比短语学研究 [J]，《外国语》(4)：32-42。
卫乃兴、陆　军，2014，《对比短语学探索》[M]。北京：外语教学与研究出版社。
翟红华、方红秀，2009，国内语义韵研究综述 [J]，《山东外语教学》(2)：8-11。
张继东、刘　萍，2006，动词happen、occur和“发生”的语言差异性探究——一项基于英汉语料库的调查与对比分析 [J]，《外语研究》(5)：19-22。

通信地址：225127 江苏省扬州市扬州大学外国语学院

基于CiteSpace的国内语料库语言学研究概述（1998–2013）*

北京外国语大学 刘 霞 许家金 北京外国语大学/燕山大学 刘 磊

提要：本文基于文献计量学工具CiteSpace对1998年至2013年发表于CSSCI刊物的语料库相关文献进行量化分析和可视化呈现。1998–2013年间，我国外语界语料库语言学的发展沿着两条主线展开，一是基于学习者语料库的中介语研究，二是基于双语语料库的翻译及对比研究。从时间上看，起初是以学习者语料库研究为主导，之后双语语料库研究更为受人关注，最近三年，国内语料库语言学研究出现了许多新的研究点，出现了跨学科、跨领域的特性。

关键词：语料库语言学、CiteSpace、中介语研究、翻译及对比研究

1. 引言

本文所关注的是国内较重要的语言学刊物登载语料库研究论文的情况。其数据来源于南京大学中国社会科学研究评价中心开发的《中文社会科学引文索引（CSSCI）数据库》。我们以“语料库”为关键词检索得到708条相关文献，截止2014年1月7日数据采集当日，708条文献的发表时间覆盖1998年至2013年初。显然，14余年，与语料库有关的语言学文献远不止于此。本文立足CSSCI数据库，集中考察高影响因子文献，应能一定程度上反映国内语料库语言学研究的大致面貌。诚然，CSSCI刊物之外不乏好文，或许今后可以基于中国知网（CNKI）的期刊数据再作进一步分析。为践行“量化、客观”原则，我们借助文献计量学分析工具CiteSpace（Chen 2006，2012；冯佳等 2014），对CSSCI中检索得到的语料库语言学文献数据做了可视化的呈现。

2. 国内语料库语言学研究文献的数据分析

2.1 1998–2013年国内语料库语言学研究总体趋势

图1所示为1998年至2013年CSSCI来源期刊登载语料库相关论文数量的分布走势。其逐年递增的总体趋势，一目了然。在数据所覆盖的14余年中，自2003年开始，似乎出现了明显的增长势头，到2010年到达顶峰，截止数据采集当日，该数据库仅收集了2013年

* 本研究得到国家社科基金项目“基于双语语料库的汉语复杂动词结构英译研究”（12CYY060）和教育部“新世纪优秀人才支持计划”（NCET-12-0790）的资助。梁茂成教授、李文中教授、熊文新副教授对本文初稿提出了细致中肯的意见和建议，特致谢忱。

第1、2期的文献。由于文献数量不全导致了2013年数量的减少。

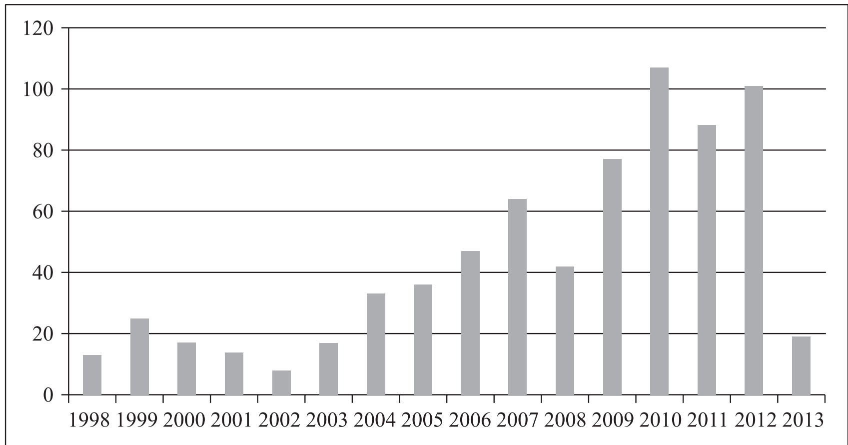


图1. 国内语料库语言学研究文献的时间分布

708篇语料库相关研究分布于不同的学科。其中最多的是语言学（586篇），其次是教育学（50篇）、图书情报学（29篇）以及新闻学与传播学（10篇）。语料库在语言学研究中的应用从1998年至今日益增多，但在其他领域的应用在近三四年间才得到推广。本研究的数据显示，教育学和图书情报学领域的第一篇语料库相关研究发表于2002年，新闻与传播学领域更晚，直到2007年才出现第一篇。但从2010年到2013年初，这三个领域应用语料库的研究几乎占据了14余年的一半，如教育学（24/50），图书情报学（13/29），新闻学与传播学（6/10）。由此可见，语料库已经越来越多地被其他领域的研究者掌握并应用到各自的研究中。

国内CSSCI期刊中，登载语料库语言学文章最多的前10个刊物及其刊文数量为：《外语电化教学》（双月）93篇，《外语教学与研究》（双月）57篇，《现代外语》（季刊）39篇，《语言文字应用》（季刊）38篇，《当代语言学》（季刊）32篇，《外语界》（双月）32篇，《外语学刊》27篇，《外语与外语教学》（双月）27篇，《外国语》（双月刊）25篇，《中国外语》（双月）23篇。语言学尤其是外国语言学类CSSCI刊物似乎都愿意接受语料库语言学研究稿件，这一定程度上也可看出各家刊物对语料库语言学这种新式研究方法的认可度。

对国内语料库语言学研究的总体趋势有了一个大致了解之后，我们将借助CiteSpace这一科学计量学的方法，进一步考察语料库语言学研究的基本情况，试图探测其发展趋势或动向，并以可视化的方式加以呈现。根据其节点类型，CiteSpace可以呈现四类可视化图谱，第一类是作者、研究机构、国别；第二类是参引文献（cited reference）之间以及被引作者（cited author）之间的被引关系；第三类是关键词和名词性术语；第四类是研究基金。本文将基于1998到2013年初的数据，分别呈现并讨论前三类可视化图谱。

2.2　国内语料库语言学研究的主要研究单位及学者

图2中的圆圈（即节点）代表研究单位及学者，其大小代表该机构或学者发表的文章数量，数量越多，节点越大。节点由不同颜色的年轮构成，每一年轮对应文章的出版时间，由内到外，年轮对应的时间由远及近。节点间的连线代表作者与作者以及作者与机构之间的联系。为了让图谱中的文字更加清晰，笔者设定阈值为4，即图2中有文字标注的节点，均是发表文章大于等于4的学者和机构。

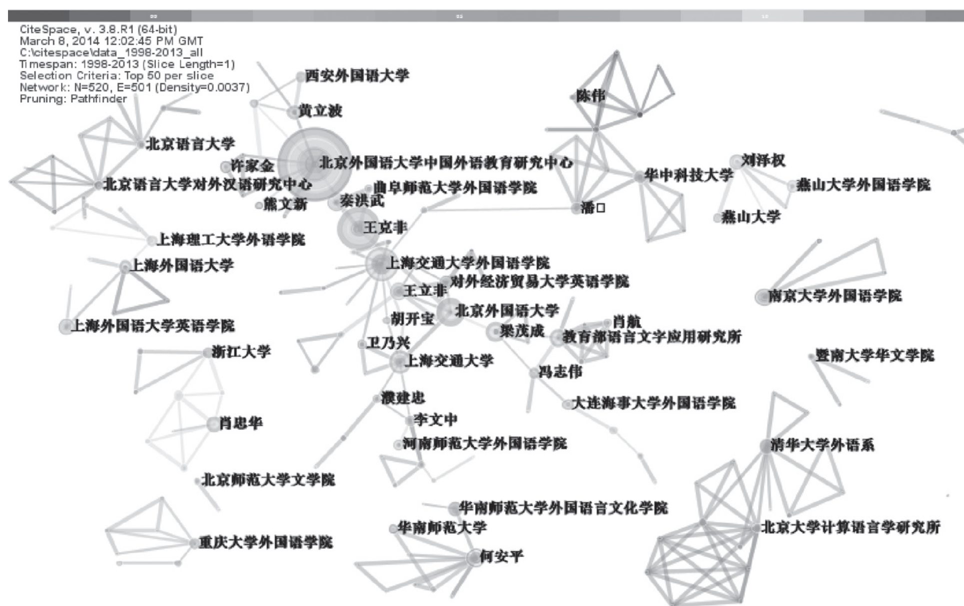


图2.　国内语料库语言学的主要研究单位及学者[1]

1998–2013年间，国内有一些产出语料库语言学研究成果较多的单位或机构，比如，北京外国语大学（及北京外国语大学中国外语教育研究中心）、上海交通大学（及上海交通大学外国语学院）、华南师范大学外国语言文化学院、上海外国语大学、南京大学外国语学院、对外经济贸易大学英语学院、清华大学外语系、华中科技大学、河南师范大学外国语学院、燕山大学外国语学院等。以上基本是外语院、校，它们从事的主要是基于语料库的英语研究。在开展语料库研究的单位中，也有一些以中文为主要研究对象的机构，如教育部语言文字应用研究所、北京语言大学对外汉语研究中心、华中师范大学语言与语言教育研究中心等。有关汉语学界基于语料库开展的研究[2]可参看Feng（2006）。

显然，图2提供的信息勾勒出了国内从事语料库语言学研究的主要机构。这些机构往往由一些核心成员形成研究团队，从而能持续产出有影响的研究成果。语料库语言学研究本质上离不开团队协作。团队的存在是语料库建设与研究成果产生的重要平台。国际范围内，兰卡斯特大学、伯明翰大学、伦敦大学学院、比利时鲁汶天主教大学、北亚利桑那大学等无不是学者聚集的团队。只有形成团队，构建学术交流机制，才能不断产生更新更好

的语料库产品和学术成果。

2.3 语料库语言学的知识结构

“一篇文献的被引频次可以在一定程度上反映该文献的影响度”（刘则渊等 2008: 143）。理清国内语料库语言学具有高被引频次的文献，以及他们的被引用情况，能够帮助我们廓清该领域的知识结构，并通过找到文献的被引激增以及转折点，发现该领域的研究动向。图3呈现了国内语料库语言学研究的共被引图谱，并标注出了国内学者引用最多的前11条文献[3]。根据节点大小，我们不难看出，被引频次从高到低依次为：Sinclair（1991）、杨惠中（2002）、Hunston（2002）、桂诗春、杨惠中（2003）、王克非等（2004）、Baker（1993）、文秋芳等（2003）、卫乃兴（2005）、黄昌宁（2002）、李文中、濮建忠（2001）、何安平（2004）。这些高频次被引文献可分为两大类：一类是关于语料库的基础知识，如Sinclair（1991）、杨惠中（2002）、黄昌宁（2002）；另一类是关于语料库的应用，在这一类中又可分为两类，一是语料库如何应用到教学和学习中，如Hunston（2002）、桂诗春、杨惠中（2003）、文秋芳等（2003）、卫乃兴（2005）、李文中、濮建中（2001）、何安平（2004），另一类是语料库如何应用到翻译研究中，如王克非等（2004）、Baker（1993）。

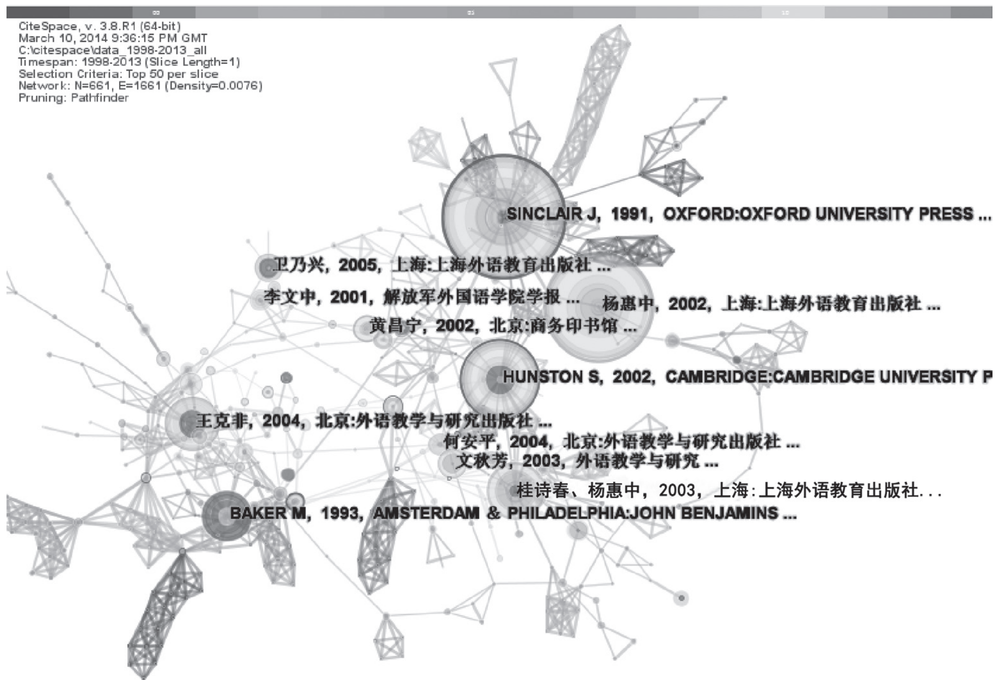


图3. 国内语料库语言学研究的共被引文献[4]

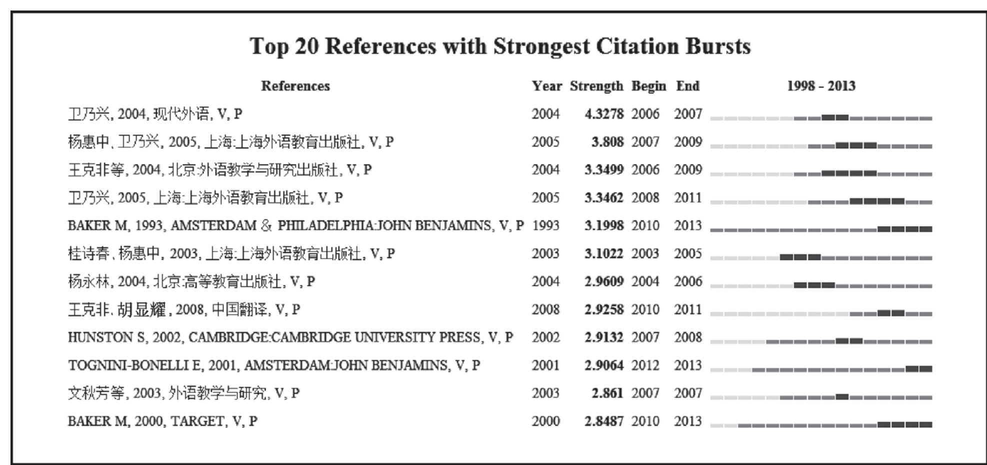
任何学科发展过程中都会经历一些具有重要意义的转折点，语料库语言学也不例外。CiteSpace通过计算每一个节点的中介中心度（betweenness centrality），继而标注出整个网

络图谱中用以连接两个子网络的节点，即转折点。表1列举了国内语料库语言学的十个转折点，并按照其中介中心度由高到低排列。

表1. 国内语料库语言学的转折点文献

	文献名称	作者	出版年代	中介中心度	总被引频次
1	Corpus, Concordance, Collocation	Sinclair	1991	0.41	67
2	Corpora in Applied Linguistics	Hunston	2002	0.33	43
3	Corpus-based Translation Studies: The Challenges That Lie Ahead?	Baker	1996	0.23	10
4	Corpus and text: Basic principles	Sinclair	2005	0.21	2
5	Corpus linguistics and translation studies: Implications and applications	Baker	1993	0.17	5
6	Corpus Linguistics	McEnery & Wilson	2001	0.16	12
7	If you look at...Lexical bundles in university teaching and textbooks	Biber *et al.*	2004	0.15	4
8	Prolegomenon to a theory of translation	Frawley	1984	0.15	2
9	The Role of corpora in investigating the linguistic behaviour of professional translators	Baker	1999	0.11	6
10	Dimensions of Register Variation: A Cross-linguistic Comparison	Biber	1995	0.11	2

此外，图3的知识图谱中还有一些在1998–2013年之间被引激增的文献。“通过对被引激增文献的考察，我们可以追踪某一学科和研究领域的热点及其历时演变”（Chen 2012：597）。在本研究的数据中，我们探测到12个被引激增的文献，如图4所示，他们均是在过去13余年间语料库语言学研究的热点。最近三年出现引用激增的文献有Baker（1993）、Baker（2000）以及Tognini-Bonelli（2001）。前两条文献足以说明将语料库应用到翻译研究是当下的研究热点，后者则体现了近期语料库语言学领域中关于两种研究范式的争论，正是Tognini-Bonelli（2001）首次对基于语料库和语料库驱动的研究范式进行了区分。通过进一步观察这三条文献的施引文献（citing reference），甚至进一步锁定引文内容，便可证实我们的猜测。但如果想要更加直观地了解语料库语言学的研究热点，还需要通过节点类型为关键词（keyword）的知识图谱来呈现。



Top 20 References with Strongest Citation Bursts

References	Year	Strength	Begin	End	1998 - 2013
卫乃兴, 2004, 现代外语, V, P	2004	**4.3278**	2006	2007	
杨惠中, 卫乃兴, 2005, 上海:上海外语教育出版社, V, P	2005	**3.808**	2007	2009	
王克非等, 2004, 北京:外语教学与研究出版社, V, P	2004	**3.3499**	2006	2009	
卫乃兴, 2005, 上海:上海外语教育出版社, V, P	2005	**3.3462**	2008	2011	
BAKER M, 1993, AMSTERDAM & PHILADELPHIA:JOHN BENJAMINS, V, P	1993	**3.1998**	2010	2013	
桂诗春, 杨惠中, 2003, 上海:上海外语教育出版社, V, P	2003	**3.1022**	2003	2005	
杨永林, 2004, 北京:高等教育出版社, V, P	2004	**2.9609**	2004	2006	
王克非, 胡显耀, 2008, 中国翻译, V, P	2008	**2.9258**	2010	2011	
HUNSTON S, 2002, CAMBRIDGE:CAMBRIDGE UNIVERSITY PRESS, V, P	2002	**2.9132**	2007	2008	
TOGNINI-BONELLI E, 2001, AMSTERDAM:JOHN BENJAMINS, V, P	2001	**2.9064**	2012	2013	
文秋芳等, 2003, 外语教学与研究, V, P	2003	**2.861**	2007	2007	
BAKER M, 2000, TARGET, V, P	2000	**2.8487**	2010	2013	

图4.　国内语料库语言学研究被引激增的文献

2.4　国内语料库语言学研究热点

关键词是每一篇文章核心内容的浓缩，如果某一关键词在该领域的文献中重复出现，就可以被视为研究热点，再将它们与所出版的时间相联系，就能发现某个领域在特定时期的研究热点。

图5直观地列出了国内语料库语言学研究的部分关键词，如果将出现频率最高的前50个关键词进行合并和分类，我们可以大致将国内语料库语言学的研究热点分为三个研究取向：中介语研究（如“外语教学”、“学习者语料库”等）、翻译研究（如“平行语料库”、“翻译共性”等）和语料库语言学的研究议题（如“语义韵”、“搭配”等），前两者的出现频率远远高于后者，而且在相当长时期内，中介语研究和翻译研究都会是我国外语界语料库语言学研究的两个主要方面。

图5中的信息还可以按年份来观察。图中左上方以翻译为主的节点（“平行语料库”、“语料库翻译学”等），揭示出近期国内语料库研究的主要关切点，而右下部分的关键词（“英语教学”、“学习者语料库”等）表明这些研究点势头趋缓。研究热点的转变反映了我国外语界语料库语言学研究的两大方面，即学习者语料库研究和双语语料库研究。学习者语料库研究的巅峰时期在过去的5–8年以前，最近五年左右，双语语料库成为国内语料库语言学研究的一个新的热点。另外，语料库语言学在中国发展至今，有关词块、搭配和意义单位的探讨在1998年至2011年期间一直为大家所关注，这与国际上语料库语言学研究趋势基本吻合，即有关短语学（phraseology）的探究始终是热门话题。其中有关意义单位的讨论出现更晚一些，到最近几年才陆续产出了一批成果。

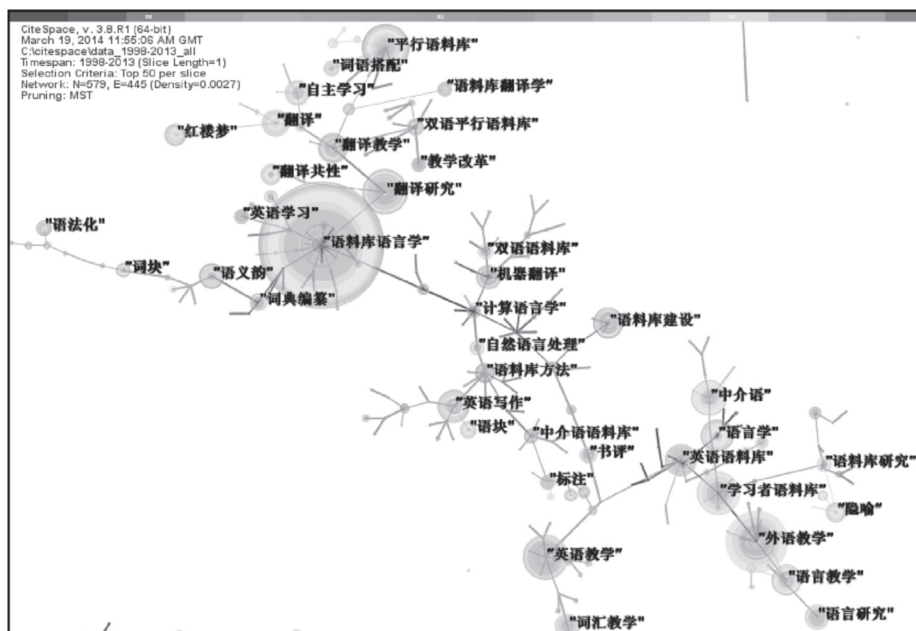


图5. 国内语料库语言学研究的关键词[5]

还有一些最近三年新出现的关键词，由于出现频率少，节点不易显现，如“口译语料库”、“汉语中介语语料库”、“web语料库”、“社会语言学”、“中国立法语言语料库”、“中文文本情感分析”、“汉语方言”、“手语语料库”、“文献计量”等，但这些关键词更能体现目前语料库语言学的研究趋势。我们不难看出，语料库逐渐成为了各研究领域的工具，越来越多的学者带着各自的理论视角投入基于语料库的研究当中，因此语料库语言学领域产生了许多新的研究点。

3. 结语

本文通过分析基于CiteSpace获得的语料库语言学研究机构和学者图谱、共被引图谱以及关键词图谱，发现1998–2013年间，我国外语界语料库语言学的发展沿着两条主线展开，一是基于学习者语料库的中介语研究，二是基于双语语料库的翻译及对比研究。从时间上看，起初是以学习者语料库研究为主导，然后双语语料库研究逐渐受到关注，最近三年，国内语料库语言学研究出现了许多新的研究点，出现了跨学科、跨领域的特性，如与文献计量学、社会语言学、汉语语言学的结合。当前国际语料库语言学研究也呈现出一些新的动向，如语料库分析与话语分析、语用分析、功能语言学研究，以及认知语言学研究结合起来，使语料库语言学的边界得到了较大拓展（参看McEnery & Hardie 2012）。我国语料库语言学研究者也应跳出学科的藩篱，重视从领域之外汲取营养，才能发掘出更多有价值的研究选题，作出更深入的研究。

本文对1998–2013年间国内语料库语言学研究所作的综述，主要基于量化数据和计算机自动分析，虽然有其自身的优势，但同时也有不足之处。一方面，CSSCI数据库本身存在一定的缺陷，所以基于该数据库的分析需要注意以下几点：第一，该数据库并未收录所有的好文，不同时间的检索结果并不一致；第二，某些文献的发表年代有误；第三，同一文献的引用格式不同（如作者的姓名和出版机构的拼写格式），这有可能影响网络节点之间的连线。对以上问题，我们都进行了详细的手工排查和批量处理，但仍不能保证没有错误。另一方面，我们需要清醒地认识到，学术评价不是单凭文献引用率可以最终裁判的。文献的高频引用更多反映的是学术关注度，而不是学术品质本身。譬如说，我们不能简单地认为，从文献引用率看，学习者语料库热度已退，因此我们就放弃了对学习者中介语的探究。相反，我们更应认真思考，如何突破中介语对比分析法的局限，探索出新的研究增长点。事实上，中介语研究仍然大有可为。同理，我们不能简单地认为某位学者被引用率高，其研究的水准一定高。文献引用有其特点，比方说，某领域的综述类文章远比某专题的实证研究更易被引用，出版早的文献较新文献引用率会偏高。

概而言之，基于文献引用的综述让我们可以更全面地了解“普罗大众”的学术关注，使我们在开展研究时，不至于闭目塞听、闭门造车。对于学术观点和研究价值的判定，新的研究选题的发掘，则在于多读、多问、多思、多行。

注释

1. 图2以一年为一个时间分区（time slice），每一时间分区内提取被引频次最高的前50篇文献的共被引数据，采用寻径算法（pathfinder），对单个时间切片和复合时间切片进行剪裁后生成知识图谱。

2. 本文基于CSSCI期刊的数据分析，未对汉语语言学和外语语言学做学科区分。从发文数量看，近10年左右国内外语界明显多于汉语界。然而，汉语界有一些重要团队、重要学者做出了许多出色的语料库研究成果。譬如，国家语言资源监测与研究中心的一些分中心：北京语言大学纸媒语料库杨尔弘教授团队、中国传媒大学有声媒体语料库侯敏教授团队、暨南大学海外华语语料库郭熙教授团队、厦门大学教材语料苏新春教授团队、华中师范大学网络媒体何婷婷教授团队、中央民族大学的民族语言赵小兵教授团队，以及南京师范大学董志翘教授、陈小荷教授的国家社科重大招标项目古汉语语料库等。更早的还包括陈小荷教授20世纪90年代初在北京语言大学开展的“八五”教委规划项目用于留学生汉语教学的“汉语中介语语料库系统”。中文方面理论研究包括张普教授的动态流通语料库理论以及在此基础上的语言资源监测，此外还有服务对外汉语教学的北京大学袁毓林教授团队的国家社科重大项目等。

3. 阈值选为14，即被引频次大于等于14的文献节点才出现标注文字。

4. 图3以一年为一个时间分区，每一时间分区内提取被引频次最高的前50篇文献的共被引数据，采用寻径算法，对复合时间切片进行剪裁后生成知识图谱。

5. 图5以一年为一个时间分区，每一时间分区内提取被引频次最高的前50篇文献共被引数据，采用最小生成树（minimum spanning tree），对复合时间切片进行剪裁后生成知识图谱。另外，为保证其他节点更加清晰，我们隐去了图谱中“语料库”这个最大节点。

参考文献

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications [A]. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.). *Text and Technology: In Honour of John Sinclair* [C]. Philadelphia: John Benjamins. 233-250.

Chen, C. 2006. CiteSpace: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. *Journal of the American Society for Information Science and Technology* 57(3): 359-377.

Chen, C. 2012. Predictive effects of structural variation on citation counts [J]. *Journal of the American Society for Information Science and Technology* 63(3): 431-49.

Feng, Z. 2006. Evolution and present situation of corpus research in China [J]. *International Journal of Corpus Linguistics* 11(2): 173-207.

Hunston, S. 2002. *Corpora in Applied Linguistics* [M]. Cambridge: CUP.

McEnery, T. & A. Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice* [M]. Cambridge: CUP.

Sinclair J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [M]. Amsterdam: John Benjamins.

冯　佳、王克非、刘　霞，2014，近二十年国际翻译学研究动态的科学知识图谱分析 [J]，《外语电化教学》（1）：11–20。

桂诗春、杨惠中，2003，《中国学习者英语语料库》[M]。上海：上海外语教育出版社。

何安平，2004，《语料库语言学与英语教学》[M]。北京：外语教学与研究出版社。

黄昌宁，2002，《语料库语言学》[M]。北京：商务印书馆。

李文中、濮建忠，2001，语料库索引在外语教学中的应用 [J]，《解放军外国语学院学报》（2）：20-25。

刘则渊、陈　悦、侯海燕等，2008，《科学知识图谱方法与应用》[M]。北京：人民出版社。

王克非等，2004，《双语对应语料库：研制与应用》[M]。北京：外语教学与研究出版社。

卫乃兴，2005，《语料库应用研究》[M]。上海：上海外语教育出版社。

文秋芳、丁言仁、王文宇，2003，中国大学生英语书面语中的口语化倾向——高水平英语学习者语料对比分析 [J]，《外语教学与研究》（4）：268-274。

杨惠中，2002，《语料库语言学导论》[M]。上海：上海外语教育出版社。

通信地址： 100089　北京市北京外国语大学中国外语教育研究中心（刘霞、许家金）

100089　北京市北京外国语大学中国外语教育研究中心/066004　河北省秦皇岛市燕山大学外国语学院（刘磊）

人文社科学术文本俄汉平行语料库的创建与研究*

陕西师范大学 陶 源

提要：创建平行语料库是描写翻译学研究的需要，国内外已有多个双语或多语平行语料库，但鲜有俄汉平行语料库。本文阐述了俄汉平行语料库的总体结构、文本选择和构建流程。语料库建设过程中，我们除了进行俄汉语料的句级对齐、篇头添加和交集检索以外，还做了以下两项工作：1）探索了俄语词屈折形式的检索问题；2）探索了学术文本语料库的术语库自动生成问题。语料库的研制是与其研究目的紧密相关的，本文对俄汉平行语料库的预期研究问题和前景进行了展望。

关键词：俄汉平行语料库、语料库创建、术语库生成、交集检索

1. 背景和意义

本文拟阐述俄汉平行语料库（人文社科学术文本）创建的过程及以该语料库为基础的可能进行的后期研究。俄汉平行语料库，尤其是用于翻译研究的平行、类比语料库在国内尚属首创，因此在语料库设计、文本搜集和预处理、检索、篇头的添加等方面都存在着众多难点。本文以语料库创建过程中的工作实例为依托，阐述语料库建设的主要环节。

1.1 国内外现有的俄语语料库

迄今为止，俄语语料库的建设相对其他语种还比较少见，尤其是包括俄语的双语语料库更是少之又少。

俄语单语语料库是在语料库语言学的背景下得以建设的。早在2003–2004年间，俄罗斯语言学界就开始了对语料库语言学的研究，俄罗斯科学院语言学研究所以Захаров为首的学者就开始创办了语料库语言学论坛，2006年又改为“语料库与计算语言学论坛”（семинар по корпусной и компьютерной лингвистике）。在历届论坛上Митрофанова、Мухин、Паничева和Ягунова从句法、语义、逻辑等角度对语料库语言学进行了论述，单语语料库也相继建成，如俄语国家语料库（Национальный корпус Русского языка），谢尔盖·沙洛夫语料库（Русские корпуса Сергея Шарова）等，其中俄语国家语料库收录3.4亿多词次，

* 本研究得到2013国家社科基金项目“基于俄汉平行语料库的人文社科类学术文本翻译研究”（13BYY026）、2013年度中央高校基本科研业务费专项资金项目“翻译过程与译者的心理选择、顺应机制研究——俄汉模糊语言翻译”(13SZYB10)的资助。

谢尔盖·沙洛夫语料库收录约4千万词次，这两个语料库都是俄语单语的通用语料库，包括文学、政论、科技等各方面内容。2008年Л.Н.Беляева在该论坛的报告中提出了语料库与翻译的关系，并对平行语料库的建设和可以运用平行语料库进行的研究内容进行了介绍和讲述。但是翻译语料库甚至双语语料库的建设还相对滞后，已经建成的译学语料库还没有见到。

国内的俄语单语库仅见：解放军外国语学院创建的新闻政论语体俄语语料库，语料库达到5,000万词次，但是该语料库是以编写网页爬虫的方式建成的，没有进行细致的分类和加工。

现有的俄汉双语库主要有：解放军外国语学院建设的俄汉翻译平行语料库（崔卫、张岚 2014）、北京第二外国语学院翻译研究中心开发研制的全国公示语翻译语料库、俄汉双语库和吕红周（2007）自建的双语库、秦立东（2007）自建的双语库等。

综观现有的俄语语料库，尤其是包含俄语的翻译语料库或平行语料库的建设还很不够，唯一具有一定规模的汉俄公示语翻译语料库也还存在着以下问题：

1）规模较小。北京第二外国语翻译研究中心原有的研制目标是一个多语种公示语平行语料库，现因为技术手段、文本选择等因素，汉俄部分仅进行了初期建设后就已经搁置了。这当然是和公示语本身的图片多、文字少，词汇多、句子少，专有名词多，缩略语多等特点是紧密相关的，但是翻译研究、翻译教学和翻译实践都需要建设更大规模的俄汉平行语料库。解放军外国语学院建设的俄汉翻译平行语料库还处于建设阶段，建成的只有军事外宣子库。

2）检索受限。公示语语料库仅能提供汉俄单向检索，俄语中很多复杂的形态变化其检索问题还没有真正解决。解放军外国语学院语料库（崔卫、张岚 2014）对于俄语部分的检索问题没有详细阐述。

本文拟建设的人文社科类学术文本俄汉平行语料库和汉俄公示语翻译语料库，都属于包含俄语的专门性双语语料库。但是我们在库容、检索和用途三个方面都有所突破：人文社科类学术文本俄汉平行语料库一期建设拟收词500万，后期将达到1,000万；实现俄语的正则表达式检索和网页检索；并使语料库能够用于语料库翻译学和俄汉语对比研究、翻译教学和学术论文写作等方面。

1.2　俄汉平行语料库创建的意义

语料库翻译学从产生到发展已经20年了，基于语料库的翻译研究方法已被应用到越来越多的研究之中。

俄汉翻译研究是国内翻译研究的重要领域，而目前现有的俄汉翻译研究还停留在对俄罗斯译学的述评（杨仕章 2001a，2001b，2002，2003）、翻译批评（朱达秋 2011，2013）和基于少量译例的翻译实践研究（陶源 2011）上，也就是说，现有的俄汉、汉俄翻译研究主要还属于规定性译学的范畴。这种路径下的俄汉、汉俄翻译研究中的译例选择主要根据研究者既有的理论前提而定，更倾向于内省式的语料选取方式；大部分论文或专著的译例都选自有限几部作品的原文和译文，而且其中大部分是文学作品；在对译例的分析上也局限于某一种理论，在微观的语料分析之后的宏观分析受到限制和制约，这种翻译研究存在

着语料少、主观性强等缺陷。总之，现有的俄汉、汉俄翻译属于规定性译学的研究范畴。

因此，把俄汉翻译研究纳入描写译学研究的范畴有其必要性，而描写性译学研究的基础就是语料库，研制俄汉平行语料库的意义如下。

1.2.1 理论意义

研制俄汉平行语料库的理论意义有两点。

1）俄汉两种语言形态方法的差异最为明显（马庆株 2002；赵敏善 1994），俄语的形态变化特别丰富：在词的层面，每一种名词都有12种变化形式，动词有时、体、态的变化，更有其他语言鲜有的形动词和副动词，其中形动词在学术文本中出现频率更是高于一般文本，而这些语言特征正是汉语中所没有的。汉语的语言特征正好与此相反，我们常见学者对汉语有这样的论述："汉语重意合"，而当今学者们在汉语研究中更是提出了"语义中心论、语义决定语法"的观点（陆俭明 2004；赵世举 2007）。把这两种语言放在一个平行语料库中，研究从俄语翻译而来的翻译语言的特征必定更具有代表性和典型性。

2）双语语料库可以分为翻译语料库、平行语料库（对应语料库）和类比语料库（王克非 2012：10），也可以分为平衡语料库（通用语料库）和专门语料库。本文要讨论的人文社科类学术文本俄汉平行语料库属于专门语料库的范畴，专门语料库指的是"关于特定主题文本的集合"，且这些文本"均由行内专家为不同的读者群所写"，这些读者群可能是同行的专家，或是缺乏相关专业知识的大众，亦可能是学生等（Kübler 2003：29, 转引自李德超、王克非 2010）。综观目前的双语语料库建设，专门应用于翻译研究的专门语料库还相对缺乏，现有的专门语料库也一般规模较小，从几万到几十万词次不等。香港理工大学研制的"新型双语旅游语料库"在第一阶段建设中已经达到200万词次，是一个集研究和应用于一体的双语专门语料库。本文要建设的也是一个专门语料库，具有不同于以往双语专门语料库的特征：（1）以俄汉语言为对象；（2）只选取人文社科类学术文本为语料；（3）建设规模将达到500万词次，且为开放型语料库，后期还将不断增加和扩容。

因此本文拟讨论的语料库具有专门性、涉及俄汉两种语言和较大规模的特点，它在建设和研制过程中必然会面临语料选取、识别、对齐、工具和检索等方面的问题和困难，而这些问题的解决也必将为后期俄汉平行语料库、学术文本语料库建设提供良好的范式。语料库的建成使国内翻译研究的路径得以拓宽，描写性译学从英汉、日汉翻译延展到俄汉翻译领域，专门性文本的描写性翻译研究将涉及一个新的领域：学术文本翻译。

1.2.2 实践意义

研制和建设人文社科类学术文本俄汉平行语料库的目的是以它为基础的俄汉翻译研究，特别是人文社科类学术文本的翻译研究，因此语料库建设以翻译实践研究和翻译教学为目标。基于该语料库我们将在下列研究和教学领域开展工作：

1）人文社科类学术文本的俄汉翻译研究包括：学术文本的翻译语言特征研究、人文社科类学术文本的术语翻译研究、人文社科类学术文本中元话语信息和某些句式的翻译研究及其表现的翻译选择问题等。

2）翻译教学，本语料库将为俄汉翻译教学提供一种新的范式，即数据驱动型教学

（data-driven learning，缩写为DDL），为这种教学模式提供较大规模的真实语料和教学资源。DDL这里指的是“在课堂上利用电子计算机生成索引（concordances）来帮助学生发掘目标语型式（pattern）中的规律，并根据索引结果来研发种种学习活动及设计学习练习”（Johns & King 1991：iii）。以本语料库为基础教师在教学中开发各种新的手段来进行俄汉、汉俄翻译教学，学生也可以利用语料库进行俄汉翻译学习，检测自己的翻译作业。

3）指导学生进行俄语毕业论文写作。本语料库选取俄语语言文学的三个研究方向，即语言学、翻译学和文学的相关论文和专著作为语料，且所有语料都采取双语对应的形式。学生在论文写作时可以通过检索平台获得各研究方向的论文结构、句式结构和常用词、词组等重要信息，也可以通过对照语料库中的相关论文，检验自己论文的单词拼写、语法错误和逻辑思维等。

2. 人文社科类学术文本俄汉平行语料库的设计和研制

本文在这一小节将重点描述人文社科类学术文本俄汉平行语料库的设计和研制思路。这是国内首个较大规模的俄汉平行语料库，我们将对它的总体设计、文本选取和抽样、对齐和检索等问题进行较为详细的介绍和描述。

语料库的建设是为后期的研究和教学服务的，研究目标决定语料库建设和研制的方方面面。人文社科类学术文本俄汉平行语料库的研究目标已在本文第一小节进行了陈述，根据翻译研究、翻译教学和指导学生论文写作的基本目标，语料库将进行下列的设计和研制。

2.1 俄汉平行语料库的设计：总体结构和文本选择

本语料库属于专用（specialized）、同质的（homogeneous）语料库，只收集原文为俄语的人文社科类学术文本及其译本，之所以选择这类文本作为语料来源，是基于以下考虑。

1）从俄汉翻译实践的状况来看，20世纪译介最多的当属文学作品，而进入本世纪，俄汉文学翻译相对于学术翻译发展滞后。本语料库收录的语言学、政治学和翻译学的译文都产生于本世纪。2）从文本的语言风格来看，学术文本的语言具有简洁、客观、概括的特点，在词和词组的使用上，学术文本多使用术语；在句式使用上，多使用某些特定句式和复合句。简洁的学术文本原文语言在译文中是否会呈现显化的特征？术语翻译是否会呈现范化的特征？高频使用的句式和复合句呈现什么样的翻译策略和转换机制？所有这些问题都将在语料库翻译学的框架下得以回答。

如上文所述，学术文本语料库是同质的语料库，它的同质性又决定了语料库在文本采样和选择方面有自己的要求，它不能像异质语料库那样进行随机性的采样，而是根据语料库的研制目的，对文本进行筛选和分类。

2.2 语料库的构成和规模

本语料库整体上由一个双语学术文本翻译对应语料库（parallel corpus，缩写为PC）与一个汉语学术文本翻译类比语料库（comparable corpora，缩写为CC）组成。在研究的第一阶段，第一个语料库暂定为各400万字／词，第二个语料库暂定为100万字／词（为统计方便，中文部分按字数计算，俄语部分按词数计算），目前各个语料库已经完成三分之一。按照我们的设计，PC与CC均为动态语料库（dynamic corpora），即在将来会按每两年一次的频率进行补充和更新，标注和加工深度也将随着研究的不断深入而加深。

PC部分是俄汉单向语料库，收录俄罗斯人文社科类专著及其对应的汉语译本，初期拟收录政治学、语言学、翻译学、文学专著共10部及其汉译本。语料库为共时库，所收录的原文均为现代俄语标准语的版本，译文也都译于20世纪80年代以后。由于俄语文本基本上没有电子版，项目组采取的是扫描、校对的方法，扫描得出的文本错误较多，要耗费大量的人力进行人工校对。

CC部分主要选取的是PC库中译者所写的学术专著，所选文本也涉及政治学、语言学、翻译学和文学四大领域，这样的设计理念是为了使PC的译文库和CC部分在以下几个方面具有可比性：1）出自相同的译者（作者），他们的语言特征具有可比性；2）均为学术文本，他们的语体特点具有可比性；3）均为全文收录，他们的上下文具有可比性；4）均为共时作品，语言的时代风格具有可比性。具体的语料库构成如图1。

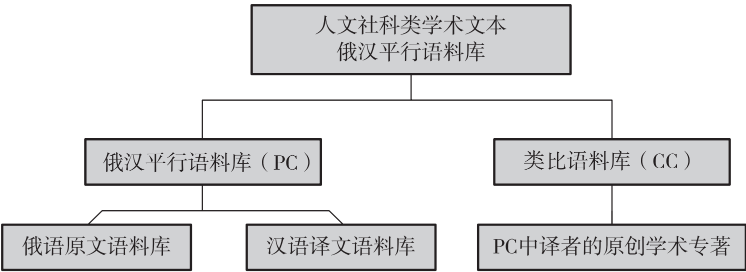


图1. 人文社科类学术文本俄汉平行语料库基本构成图

PC中的翻译语言可以和俄语原文语料库中的语言进行对比，也可以和CC中的同一领域、同一译者（作者）的语言进行对比。PC中每一本专著的原文和译文形成一个子库，CC中的每一本原创学术文本也形成一个子库，这样就形成了PC1、PC2…… CC1、CC2…… PC1中的词、词组可以在子库内进行俄汉语的对比，也可以在整个PC中进行分析，还可以通过和CC1的对比进行定量研究。

为使后期的翻译研究更能代表整个人文社科类学术文本翻译的全貌，语料库在收录和采用环节就尽量注意各类文本的平衡，目前完成的多为政治学、语言学文本，在后期的建设中将注意增加文学和翻译学文本的采样数，以期实现各类文本的基本平衡。

语料库收录文本涉及的人文社科4个研究领域的具体采样书目和词数如表1。

表1. 俄汉学术文本平行语料库语料采集情况表

	平行语料库			类比语料库	
	俄语	汉语字	样本数	汉语字	样本数
政治学和国际关系	418,100	710,856	4	303,718	2
语言学	568,738	855,326	3	335,546	2
文学	208,643	315,258	5	316,328	2
翻译学	401,223	598,816	2	287,914	2
合计	1,596,704	2,480,256	14	1,243,496	8

2.3 俄汉平行语料库篇头信息的设计与添加

在俄汉平行语料库的篇头我们拟涉及如下主要信息：语言（language）、种类(type)、作者（author）、译者（translator）、年份（time）和标题（title）共六项篇头信息。其中语言包括俄语和汉语，种类包括政治学和国际关系、语言学、文学和翻译学，年份包括创作年份和翻译年份，标题可以是俄语或汉语。

设计与添加语料库篇头信息的目的是实现语料的网上检索功能，即使用者能够根据这些信息检索到所需要的文本，因此语料库研制过程中我们采用创建属性值表→导入数据库→添加属性值的方法来完成篇头信息的添加过程。具体的添加过程如下。

1）编写属性，属性值为：俄语（ru），中文（ch），种类（type），作者（author），译者（translator），年份（time），标题（title）（见图2）。

2）导入俄汉平行语料库：导入平行语料库，通过java语言在eclipse上实现已对齐的文本文档txt逐句自动导入mysql数据库中（其中遇到数据库无法识别的字符就会报错，然后手动粘贴报错的那一行，再次运行则可继续导入，实现了比较快速的导入数据库）。由于每一个分库的语料取自一本书，该分库的后五个属性（种类，作者，译者，年份，标题）都是同一值，因此可批量写入（见图3）。



Table: lan

字段信息

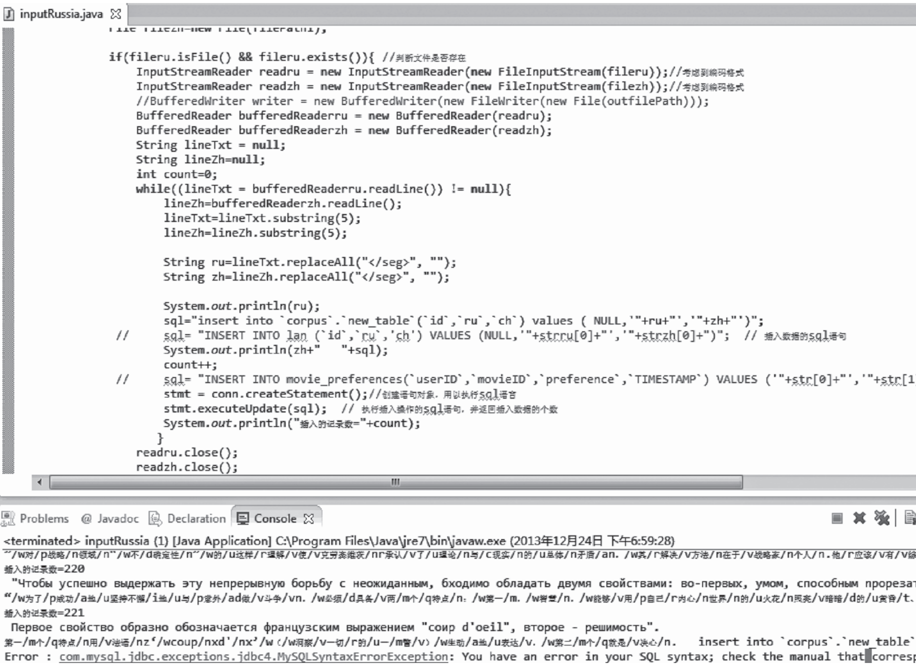
Field	Type	Comment
id	int(11)	
ru	varchar(8000)	
ch	varchar(8000)	
type	varchar(20)	
author	varchar(20)	
translator	varchar(20)	
time	varchar(20)	
title	varchar(20)	

图2. 属性值截图

3）在数据库中写入代码，添加属性值：编写导入语句的代码，具体代码见图4。通过jdbc连接数据库后按行插入，由于定义了count++，实现导入一条后自增，插入失败后报错能发现是哪一行无法插入数据库（由于数据的格式或特殊字符原因导致无法自动插入），如图4所示导入成功至221行，则手工粘贴222行进数据库，再继续运行本语句可继续导入。如此反复几遍，整个数据库导入和属性值添加可完成，见图4。

	id	ru	ch	type	author	translator	time	title
	1	УЧЕБНОЕ ПОСОБ	前言/n本书/r旨在/v阐述	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	2	В настоящее в	目前/t，/w至少/d在/p三	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	3	Жестких грани	在/p这儿/r不/d存在/v	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	4	К примеру, див	比如/v，/w“/w思维/n工	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	5	В литературе	在/p21/m世纪/n初/f关于	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	6	Характерный п	里/f卡尔德斯/nK的/u	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	7	Бизнес -- не во	商业/n不/d是/v战争/n，	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	8	Не случайно д	战略/n经营/vn管理/vn	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	9	Хотя Бойд и ег	尽管w伯伊德/nr及其/c追	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	10	Считается, что	“对于/p所有/b的/u经理	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	11	Вместе с тем о	同时/c显而易见/i的/u是	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	12	Что касается	至于/p谈到/v在/p分析/	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	13	Шродт, наприме	比如说/l，/w施罗德/nr	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	14	И экономическ	作为/p重要/a成分/n，/w	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	15	Лучшим пример	游戏/n理论/n可以/v作为	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	16	Хотя политоло	尽管/c现在/t政治学/n从	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	17	Вместе с тем не	同时/c必须/d明白/v并/	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
	18	В качестве пр	比如说/l，/w美国/ns议	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)

图3.　对齐语句导入数据库截图



```
inputRussia.java

        if(fileru.isFile() && fileru.exists()){ //判断文件是否存在
            InputStreamReader readru = new InputStreamReader(new FileInputStream(fileru));//考虑到编码格式
            InputStreamReader readzh = new InputStreamReader(new FileInputStream(filezh));//考虑到编码格式
            //BufferedWriter writer = new BufferedWriter(new FileWriter(new File(outfilePath)));
            BufferedReader bufferedReaderru = new BufferedReader(readru);
            BufferedReader bufferedReaderzh = new BufferedReader(readzh);
            String lineTxt = null;
            String lineZh=null;
            int count=0;
            while((lineTxt = bufferedReaderru.readLine()) != null){
                lineZh=bufferedReaderzh.readLine();
                lineTxt=lineTxt.substring(5);
                lineZh=lineZh.substring(5);

                String ru=lineTxt.replaceAll("</seg>", "");
                String zh=lineZh.replaceAll("</seg>", "");

                System.out.println(ru);
                sql="insert into `corpus`.`new_table`(`id`,`ru`,`ch`) values ( NULL,'"+ru+"','"+zh+"')";
 //             sql= "INSERT INTO lan (`id`,`ru`,'ch') VALUES (NULL,'"+strru[0]+"','"+strzh[0]+")";  // 插入数据的sql语句
                System.out.println(zh+"    "+sql);
                count++;
 //             sql= "INSERT INTO movie_preferences(`userID`,`movieID`,`preference`,`TIMESTAMP`) VALUES ('"+str[0]+"','"+str[1
                stmt = conn.createStatement();//创建语句对象，用以执行sql语言
                stmt.executeUpdate(sql);  // 执行插入操作的sql语句，并返回插入数据的个数
                System.out.println("插入的记录数="+count);
              }
            readru.close();
            readzh.close();

Problems  @ Javadoc  Declaration  Console
<terminated> inputRussia (1) [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (2013年12月24日 下午6:59:28)
”/w对/p战略/n领域/n”/w不/d确定性/n”/w的/u这样/r理解/v使/v克劳塞维茨/nr承认/v了/u理论/n与/c现实/n的/u总体/n矛盾/an. /w其/r解决/v方法/n在于/v战略家/n个人/n.他/r应该/v有/v
插入的记录数=220
 "Чтобы успешно выдержать эту непрерывную борьбу с неожиданным, бходимо обладать двумя свойствами: во-первых, умом, способным прорезат
“/w为了/p成功/a地/u坚持不懈/i地/u与/p意外/ad做/v斗争/vn. /w必须/d具备/v两/m个/q特点/n: /w第一/m. /w智慧/n. /w能够/v用/p自己/r内心/n世界/n的/u火花/n照亮/v暗暗/d的/u黄昏/t、
插入的记录数=221
 Первое свойство образно обозначается французским выражением "coup d'oeil", второе - решимость".
第一/m个/q特点/n用/v法语/nz‘/wcoup/nxd'/nx’/w (/w洞察/v一切/r的/u一/m瞥/v) /w生动/a地/u表达/v. /w第二/m个/q就是/v决心/n.    insert into `corpus`.`new_table`
Error : com.mysql.jdbc.exceptions.jdbc4.MySQLSyntaxErrorException: You have an error in your SQL syntax; check the manual that corresp
```

图4.　导入语句代码截图

全部导入3,086行并添加属性值后的数据库见图5。

	id	ru	ch	type	author	translator	time	title
□	3073	Форматы "плани	规划/n"的/u大小/n在/p	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3074	В данном пособ	本书/r中/j政治/n战略/r	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3075	Стратегически	战略/n领袖/n应该/v特别	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3076	Здесь получае	在/p这里/r战略/n思想/r	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3077	Стратегически	战略/n分析/v可以/v为/p	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3078	Стремясь к аде	战略/n分析/vn希望/n完	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3079	Рутинное план	墨守陈规/i的/u规划/n会	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3080	Какой именно и	不/d确定/v"结局/n"中/f	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3081	В силу комплек	由于/p这些/r元素/n的/u	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3082	Различные мето	分析/v的/u不同/a方法/r	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3083	Обобщая, можно	总之/c，/w可以/v说/v政	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3084	С одной сторон	一方面/c，/w政治/n过程	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3085	С другой -- это	另一方面/c，/w资源/n/	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
□	3086	Стратегически	战略/n分析/vn"计算/v出	政治学	奥日加诺夫	胡谷明	2007	政治战略分析
*	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)

图5.　添加属性值后数据库截图

2.4　俄汉平行语料库的检索

人文社科学术文本俄汉平行语料库的检索和术语库生成是语料库研制中的重点和难点。汉语没有词形变化，可以直接用“%+汉语词+%”实现。它可以在Word页面和ParaConc页面下辅助检索。

俄语属于屈折语，俄语中每一个词的曲折形式可以达到10几种之多，如который在我们的学术文本平行语料库中属于高频词，而该词的屈折形式有：который, которое, которая, которые, которого, которой, которому, которым, котором, которых, которыми, которых等，用简单的正则表达式“%...%”检索很难一次检索到文本中所有的*который*，而在ParaConc上检索只能使用该软件允许的4种通配符，即：%，*，?和@。针对*который*一类词，运用ParaConc允许的通配符，编写专门通配符检索，这类通配符和俄语字母组成的表达式也属于正则表达式的范畴，如正则表达式：<[Кк]отор*>可以一次实现对上述который十一种形式的检索，在Word中检索<[Кк]отор*>得到图6页面：

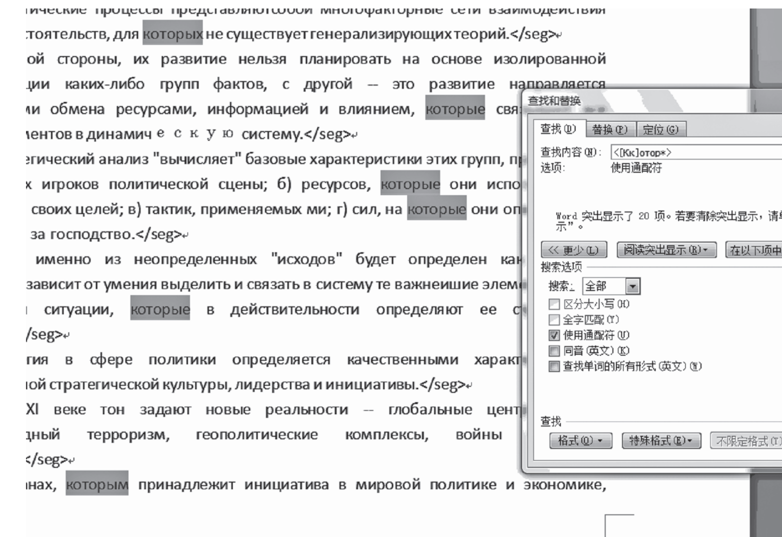


图6.　运用通配符在Word页面下检索截图

而本语料库的研制目的是用于翻译研究，所以更多地需要实现在ParaConc和AntConc等软件支持下的检索，正则表达式<[Кк]отор*>在ParaConc软件支持下对语料库的政治学子库中一部分样本进行检索后得到图7页面。



图7. ParaConc页面下**который**的通配符检索截图

除了支持在Word、ParaConc、AntConc等程序中的检索，本项目还将为俄汉学术文本平行语料库创建专业的网上检索平台。该检索平台除了可以实现词汇和短语的检索外，还将实现2.3中5项属性值的交集检索，即可以允许检索同时符合1–5个属性值的句子或文本。这一功能的实现第一步就是2.3中所述的导入数据库，在数据库中编写代码，具体代码截图如图8。



图8. 交集检索代码截图

在俄语和中文的切换查找中，if（false）查找俄语，而把括号中的false改成true后则是查找中文。需手动更改括号中的true或false属性来进行切换。到后期，在网页上实现时就可以变成俄语查找和汉语查找的选项，功能与有道词典相似。

我们试着查找俄语стратегия一词的结果，属性值为：种类：政治学；作者：奥日加诺夫；译者：胡谷明；年份：2007；标题：政治战略分析（见图9）。

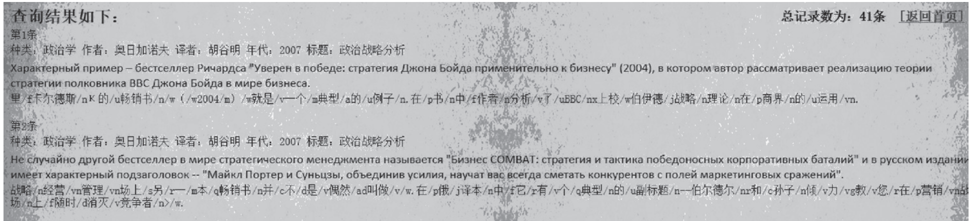


图9.　**стратегия**与5种属性值交集检索示例图

图8中光标所在的那一行（String Chinese=“%政治%”）下数7行（到String sqltitle=“%政治战略分析%”）。其中两个%之间分别填入所要查找的中文或者俄语，后面填入想要限定的种类：作者、译者、年份、标题。如果要进行相应的模糊检索，即检索一个词的多种变化形式，可以同时使用3.5中所述的通配符和本小节的交集检索。

2.5　俄汉学术文本语料库的术语库生成

本项目的研究对象是人文社科类学术文本，其中包括政治学与国际关系、语言学、文学和翻译学四个方向的专著，原文和译文中含有大量一一对应的社科类术语。术语翻译也就成为后期翻译研究的重要课题之一，这也是学术文本语料库和通用语料库及其他专门语料库的主要不同之处。因此本项目在语料库研制阶段把术语库的自动生成作为主要技术目标之一。

我们把对齐后的双语文本导入数据库，生成俄汉人文社科类学术文本平行语料库的术语库，在这一术语库中可以对俄汉两种语言的术语进行检索。下面是政治学子库的一个分库导入数据库和检索的情况。

该分库共计3,086行，导入数据库后得到以下页面（见图10）。

	id	ru	ch
□	3075	Стратегический лидер	战略/n领袖/n应该/v特别/d渴望/v看/v清/a事情/n
□	3076	Здесь получает свои п	在/p这里/r战略/n思想/n有/v了/u自己/r的/u权利/n，
□	3077	Стратегический анализ	战略/n分析/v可以/v为/p理解/v政治/n决议/n提供/v
□	3078	Стремясь к адекватном	战略/n分析/vn希望/n完全/ad相符/v地/u完成/v这个/
□	3079	Рутинное планирование	墨守陈规/i的/u规划/n会/v找到/v数千/m条/q理由/n
□	3080	Какой именно из неопре	不/d确定/v“结局/n”中/f的/u哪/r一个/m将/d被/p评
□	3081	В силу комплексности	由于/p这些/r元素/n的/u综合性/n和/c它们/r之间/f
□	3082	Различные методы анал	分析/v的/u不同/a方法/n和/c政治/n过程/n复制/v的/
□	3083	Обобщая, можно сказать	总之/c，/w可以/v说/v政治/n过程/n是/v人/n与/c局
□	3084	С одной стороны, их ра	一方面/c，/w政治/n过程/n的/u发展/vn不/d可能/v根
□	3085	С другой -- это развити	另一方面/c，/w资源/n/信息/n/影响/v的/u交换/vn的
□	3086	Стратегический анали	战略/n分析/vn“计算/v出/v”这些/r种类/n的/u基础/n
*	(NULL)	(NULL)	(NULL)

图10.　文本导入数据库截图

2.6 俄汉学术文本语料库的网页制作

把文本导入数据库，经过初步制作，语料库的网页基本形成（见图11），网页支持上述的语料检索、交集检索和术语检索。检索**метод анализа иерархий**（等级分析法）这个术语，得到两条结果（见图12）。



图11. 语料库主页截图

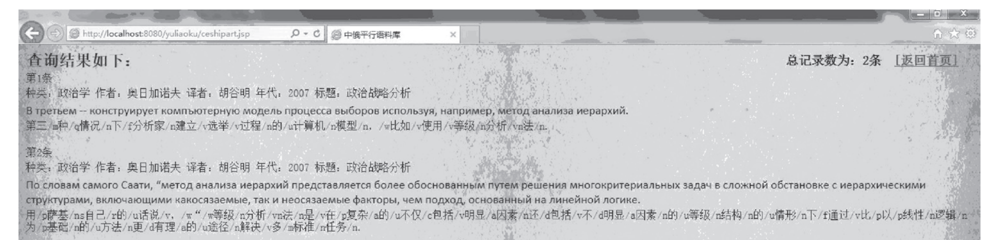


图12. 在术语库中检索**метод анализа иерархий**（等级分析法）截图

3. 俄汉平行语料库预期实现的研究内容

语料库建设的目的是翻译研究，所以在研究方法上我们首先要参阅俄汉语言对比的相关文献，从俄语语体学和学术文本的语言特色入手，寻找俄语学术文本的一些高频词、词组或句式，并对它们可能得出的译文的语言特征进行假设。基于语料库的研究最关键的一步就是以语料库的相关统计数据对这些假设进行验证。

翻译语言特征也称为翻译普遍性或翻译普遍特征，是“译文中呈现的有别于源语文本的一些典型的、跨语言的、有一定普遍性的特征”（柯飞 2005：303-307，转引自胡开宝 2011）。翻译语言作为一种客观存在的语言变体，相对于目的语原创语言，在整体上呈现出一些带有普遍规律性的特征（胡开宝 2011）。我们认为，翻译语言的特征或共性是客观存在的，而不

同语言、不同语体的文本在翻译过程中所呈现的翻译语言特征有可能具有某些微观的差异。因此，本项目把俄汉学术文本翻译中呈现的俄汉语言对比和学术文本标记语的翻译语言特征纳入微观考察的范围。根据语料库翻译学的一般方法，翻译语言特征的验证是在对比中完成的，对比包括原文和译文语言的对比、译文和原创语言的对比等。预期的研究内容包括：

1）通过俄语不定代词和不定副词在译文中的体现探讨俄译汉语学术文本的语言特征，其中包括кто-то、кто-нибудь、кое-кто、что-то、что-нибудь、кое-что когда-то、когда-нибудь、кое-когда、где-то、где-нибудь、кое-где的翻译等；

2）通过原文俄语文本中的谓语副词在译文中的体现研究汉译俄语学术文本的语言特征，其中包括можно、нужно、надо、необходимо、нельзя等；

3）通过汉译俄语学术文本中被动式（其中包括被动形动词、带ся动词等）的翻译和汉语原创文本中被动式的对比探讨翻译的语言特征；

4）通过翻译的学术文本和原创学术文本的类符/型符比、平均句长等的对比，探讨学术文本翻译语言的相关特征；

5）俄语复合句的翻译操作规范研究。这类研究以俄语中某些典型的连接词连接的复合句，如以чтобы从句为对象，着眼于以下四个规范：显化-隐化规范、简化-复杂化规范、规化-异化规范、成语化-去成语化规范。

隐化规范指原文中语法或词汇手段表达的信息在译文中隐去，取而代之的是词序、语义或上下文的手段。简化规范指原文中复杂的语言结构变成简单的语言结构，复杂化规范则相反。规化指译文与译入语规范契合度高，可读性强。异化指译文受原文影响大，更多地顺应于原文的上下文。成语化指原文自由搭配表达的信息在译文中用成语来表达。去成语化指原文的成语译为自由搭配。

在数据提取过程中，研究运用的工具有：ParaConc、WordSmith等，选取的研究对象有：（1）连接词чтобы，它的隐现和对比可以验证显化与隐化规范；（2）чтобы连接的5类从句，它们被译为简单句的比率可验证简化规范；（3）平行库中汉语部分“以便”、“以免”（чтобы的对应译文），它们与类比语料库对比可以验证规化和异化假设；（4）连接词чтобы的成语化结构，它们被译为自由搭配可验证去成语化规范。

以隐化规范研究为例，在300万词次的俄汉平行语料库政治学和语言学子库中统计得到下列结果（见表2）。

表2.　чтобы汉译隐化例证及百分比

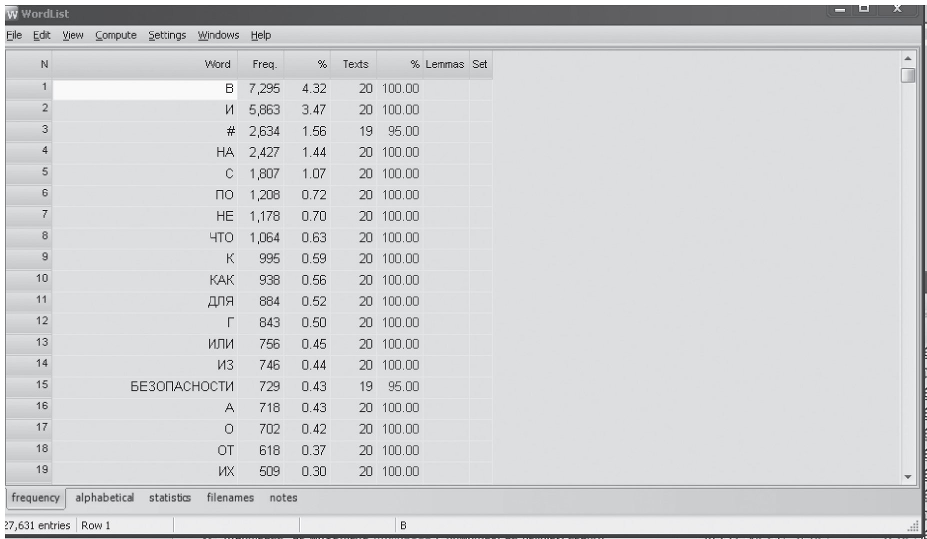
俄语词	频次	汉译	频次/频率	隐化	隐化率
чтобы	359	为了	65（18.1%）	89	24.8%
		目的是	33（9.2%）		
		要	35（9.7%）		
		以便	56（15.6%）		
		以免，免得	48（13.4%）		
		使，使得	33（9.2%）		

由表2可知，连接词чтобы在原文中共出现359次，共有6种对应译文，占所有对应译文的75.2%。语料库中чтобы有89次省去不译，隐化率为24.8%，说明чтобы从句汉译时支持隐化规范。这类从句翻译隐化规范的形成主要可能存在两个原因：一是与两种语言的特点有关。俄语和汉语在形式化程度上的差距远远超过英语和汉语。在词法上，俄语的所有实词都有十余种变化形式，且一个简单句中的各实词部分的变化是互相关联的；在句法上，俄语严格区分简单句和复合句，各复合句类型之间的界限也相对清晰，过渡区域较小（陶源 2006）。每一种复合句均有固定的连接词连接，如可以连接目的从句的连接词一般限定为：что、чтобы，其他连接词如где、когда、так...как等不能实现这一功能。语料库中原文部分的多种复合句需要чтобы连接，不能省去。而汉语的形式化程度远远低于俄语，同样是目的关系，汉语中可以不用连接词，而只通过上下文、语序或语义关系来实现，研究汉语的学者甚至提出语义决定语法的观点（赵世举 2007）。二是чтобы可以连接目的、说明、方式/程度和限定从句。其中表示目的关系时，чтобы有对应的汉语译文，而表示说明和限定关系时，其在汉语中没有固定的对应译文。

6）基于俄汉平行语料库的翻译单位研究。研究立足于语用层面，以语料库语言学的意义单位理论为基础，提出短语是翻译的基本单位，并在平行语料库中进行验证：（1）节点词本身的意义离散，要通过语境和类联接组成短语后其意义才能得以稳定，因此短语是源语中的语言使用单位；（2）短语在源语-译入语和译入语-源语的双向过程中具有语义完整性，对应率高，匹配性良好；（3）从平行语料库中抽取翻译实例，说明短语在翻译转换和对应中具有承接词项和句子的作用。

研究首先在俄语单语库中进行wordlist和concord分析，为节点词 с 的抽取奠定基础，单语库中的分析见图13。

由图13和图14可知，с 是俄语中的高频节点词，具有较强的短语趋势，短语с помощью具有典型性，可作为翻译单位研究的对象。



N	Word	Freq.	%	Texts	%	Lemmas	Set
1	В	7,295	4.32	20	100.00		
2	И	5,863	3.47	20	100.00		
3	#	2,634	1.56	19	95.00		
4	НА	2,427	1.44	20	100.00		
5	С	1,807	1.07	20	100.00		
6	ПО	1,208	0.72	20	100.00		
7	НЕ	1,178	0.70	20	100.00		
8	ЧТО	1,064	0.63	20	100.00		
9	К	995	0.59	20	100.00		
10	КАК	938	0.56	20	100.00		
11	ДЛЯ	884	0.52	20	100.00		
12	Г	843	0.50	20	100.00		
13	ИЛИ	756	0.45	20	100.00		
14	ИЗ	746	0.44	20	100.00		
15	БЕЗОПАСНОСТИ	729	0.43	19	95.00		
16	А	718	0.43	20	100.00		
17	О	702	0.42	20	100.00		
18	ОТ	618	0.37	20	100.00		
19	ИХ	509	0.30	20	100.00		

图13. 自建语料库wordlist分析

Concord

File　Edit　View　Compute　Settings　Windows　Help

N	Concordance	Set		Word #	Sent	Sent	Para	Para	Head	Head	Sect	Sect	File	
30	"управлять" иерархичес кой моделью с помощью математических средств,			6,676	321	52%	0	56%			0	56%	政治学5-3.ru.t	20
31	усложняет задачу слежения за ней с помощью загоризонтных РЛС. С			3,396	170	89%	0	44%			0	44%	政治学1-9.txt	20
32	конструкцией, обеспечивающей с помощью простых правил анализ			4,790	233	35%	0	40%			0	40%	政治学5-3.ru.t	20
33	-- каждый из них изучает свой объект с помощью специальных методов и			797	36	60%	0	7%			0	7%	政治学5-3.ru.t	20
34	объяснения сложных объектов с помощью редукционистской логики.			876	40	77%	0	7%			0	7%	政治学5-3.ru.t	20
35	информация. Объект описывается с помощью понятий: внутренняя			650	29	36%	0	5%			0	5%	政治学5-3.ru.t	20
36	первой может быть определен с помощью таких показателей, как			3,637	188	23%	0	31%			0	31%	政治学5-3.ru.t	20
37	Это вторжение, осуществленное с помощью полученного от России			7,087	303	25%	0	63%			0	63%	政治学5-6.ru.t	20
38	ячеек и организаций осуществляется с помощью шифрованных			3,438	160	37%	0	45%			0	45%	政治学1-4.txt	20
39	отношений, не можетбыть отображен с помощью ни рефлексивного			10,293	507	75%	0	87%			0	87%	政治学5-3.ru.t	20
40	приоритет фактора отражается с помощью коэффициента, для			8,105	411	32%	0	68%			0	68%	政治学5-3.ru.t	20
41	также могут быть охарактеризованы с помощью целого набора			3,823	193	56%	0	32%			0	32%	政治学5-3.ru.t	20
42	свои сектора и может быть оценен с помощью системных характеристик,			2,926	148	52%	0	25%			0	25%	政治学5-3.ru.t	20
43	решений и т#п. <seg>С помощью таких средств создается			9,678	477	8%	0	81%			0	81%	政治学5-3.ru.t	20
44	положение, образование и т# п#? С помощью кластерного анализа			6,022	284	20%	0	89%			0	89%	政治学5-5.ru.t	20
45	лидера могут быть переданы с помощью ритуальных средств его			3,099	111	12%	0	14%			0	14%	政治学5-4.ru.t	20
46	самостоятельных переменных, с помощью которых описывается и			11,468	480	72%	0	86%			0	86%	政治学5-2.ru.t	20
47	Возможность перехвата с помощью ПРО большей части			4,538	233	17%	0	59%			0	59%	政治学1-9.txt	20

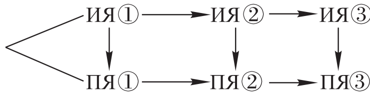
图14.　自建语料库中短语**с помощью**的concord分析

卫乃兴（2011）应用Altenberg（1999）提出的相互对应率计算词语互译的概率，对它们的对应关系进行统计。根据这一计算范式，统计с помощью的对应率，我们发现，它的源语-译入语对应率为98%（100/102），译入语-源语对应率为98.8%（79/80），以该短语为翻译单位，对应率高，因此典型短语с помощью可以作为翻译单位。

语料库中的更多翻译实例也可以验证短语作为翻译单位的假设，请看例（1）：

例：（1）В Молдове① началась полномасштабная гражданская война② со всеми характерными для таких войн событиями③. 在摩尔多瓦①开始了大规模国内战争②,具备了所有这类战争事件应有的特征③。

根据Виноградов（1977）经典短语学理论和“半预制短语”的特征（Sinclair 1996），我们把例（1）的原文和译文分别划分为3个短语，以ИЯ代表源语，以ПЯ代表译入语，两句可以形成以下的对应关系：



其中ИЯ③似乎与上文所述的с短语不符，但其实для加第二格构成的短语是с短语中名词событие的定语，因此для таких войн也是с短语的一部分，с仍然是该短语的节点词。

通过图14和例句我们看出，ИЯ①，ИЯ②和ИЯ③通过语境类联接而成源语句子，ПЯ①，ПЯ②和ПЯ③类联接而成译入语句子，三个短语“双双匹配”两句在短语对应的基础上形成更大层面的句对应。

7）除了翻译的语言特征研究，语料库还将支持翻译规范和实践研究。其中翻译规范除了操作规范以外，还将进行术语翻译规范研究，而翻译实践研究拟选择下列对象，进行源语词汇、短语和目的语词汇、短语等值性的确定，探讨俄汉语言语句之间的对应关系及

其转换规律：（1）俄语学术文本中的话语标记语在汉语中的对应词和词组，如речь идёт о、как указывалось、как отмечалось、согласно этому、в соответствии с чем、в результате、следовательно、ввиду этого、в зависимости от этого、в связи с чем、рассмотрим、перейдём к рассмотрению、итак、короче говоря、иначе говоря、из этого следует；（2）俄语中带который的说明从句的翻译转换规律；（3）俄语中带так...как的程度方式从句的翻译转换规则。

俄汉学术文本平行语料库为开放式语料库，在语料库建设阶段我们还将不断增加新的文本，扩充库容，将发现更多适合翻译研究的新问题。

4. 结语

本文在对国内外现有的翻译语料库和俄语语料库简单回顾的基础上探讨了俄汉学术文本平行语料库的研制目的、意义、文本选择和规模、面临的主要技术问题及其解决办法。语料库所选的文本涉及人文社科类学术文本的四个相关领域，四个子库在文本选择的规模上力求平衡，具有代表性。

由于已有的俄语语料库建设的技术手段还不太成熟，俄汉学术文本平行语料库在研制过程中必然面临对齐、检索、术语库生成等众多的技术难题，本文对这些难题的解决方法进行了研究，对语料库研制中已经攻克的技术问题进行了介绍。语料库的研制过程将为后期的俄汉平行语料库研制提供可行的方法和手段。

俄汉学术文本平行语料库的研制是以翻译研究为目的的，语料库的建成必将为汉译俄语学术文本的翻译语言特征研究、翻译规范研究和俄汉翻译实践研究等提供可靠的数据支持和论证理据。

参考文献

Altenberg, B. 1999. Adverbial Connectors in English and Swedish: Semantic and lexical correspondences [A]. In H. Hasselgard & S. Oksefjell (eds.). *Out of Corpora: Studies in Honour of Stig Johansson* [C]. Amsterdam: Rodopi. 249-268.

Johns, T. & P. King.1991. Classroom concordancing [J]. *ELR Journal* 4: i-v.

Kübler, N. 2003. Corpora and LSP translation [A]. In F. Zanettin, S. Bernardini & D. Stewart (eds.). *Corpora in Translator Education* [C]. Manchester：St. Jerome.

Sinclair, J. 1996. The search for units of meaning [J]. *Textu* IX: 75-106.

Виноградов, В. 1977. *Избранные труды. Лексикология и лексикография* [M]. Москва: Наука.

Беляева, Л. *Корпусная лингвистика и проблемы перевода* [OL]. (РГПУ им. А.И. Герцена). (http://corpora.phil.spbu.ru/Works2011/Беляева_86.pdf) (accessed 2014/02/16).

Захаров, В. Корпусная лингвистика как новое направление в языкознании [OL]. (http://ru.convdocs.org/docs/index-144001.html）(accessed 2014/02/16).

Митрофанова О., Мухин А., Паничева П. *Автоматическая классификация лексики в неразмеченных русскоязычных тестах* (СПБГУ) [OL]. (http://rpp.nashaucheba.ru/docs/

index-6537.html) (accessed 2014/02/16).

Ягунова, Е. *Неоднословные целостности в словаре и корпусе* [OL].（http://corpora.iling.spb.ru/seminar.htm）(accessed 2014/02/16).

崔　卫、张　岚，2014，俄汉翻译平行语料库及其应用研究 [J],《解放军外国语学院学报》（1）：81-87。

胡开宝，2011,《语料库翻译学概论》[M]。上海：上海交通大学出版社。

柯　飞，2005，翻译中的隐和显 [J],《外语教学与研究》（4）：303-307。

李德超、王克非，2010，新型双语旅游语料库的研制和应用 [J],《现代外语》（1）：46-54。

陆俭明，2004，词的具体意义对句子意思理解的影响 [J],《汉语学习》（2）：1-5。

吕红周，2007，俄汉双语语料库语义范畴自动标注[D]。硕士学位论文。哈尔滨：黑龙江大学。

马庆株，2002,《著名中年语言学家自选集・马庆株卷》[M]。合肥：安徽教育出版社。

秦立东，2007，基于俄汉熟语语料库的俄语熟语模式化及自动识别[D]。硕士学位论文。哈尔滨：黑龙江大学。

陶　源，2006，论俄语复合句分类中的模糊现象 [J],《山东外语教学》（1）：43-46。

陶　源，2011，论顺应论视角下的俄汉委婉语互译——一个新的视角 [J],《中国俄语教学》（4）：83-87。

王克非，2012,《语料库翻译学探索》[M]。上海：上海交通大学出版社。

卫乃兴，2011，基于语料库的对比短语学研究 [J],《外国语》（4）：32-42。

杨仕章，2001a，科米萨罗夫的翻译思想——《现代翻译学》评介 [J],《外语与外语教学》（4）：51-53。

杨仕章，2001b，苏、俄翻译理论中等值思想的演变 [J],《中国俄语教学》（1）：46-52。

杨仕章，2002，俄语现代翻译理论之概述 [J],《解放军外国语学院学报》（6）：62-65。

杨仕章，2003，科米萨罗夫翻译思想管窥 [J],《中国俄语教学》（3）：45-50。

赵敏善，1994,《俄汉语对比研究》[M]。上海：上海译文出版社。

赵世举，2007，试论词汇语义对语法的决定作用 [J],《武汉大学学报》（2）：173-179。

朱达秋，2011，谈学术著作翻译的常态性批评——兼评别尔嘉耶夫的《俄罗斯思想》中文译本 [J],《中国俄语教学》（1）：1-6。

朱达秋，2013，再谈学术著作翻译的常态性批评——以《俄罗斯思想》的中文译本为例 [J],《中国俄语教学》（1）：1-6。

通信地址：710062 陕西省西安市陕西师范大学外国语学院

语体文应用字汇

国立东南大学 陈鹤琴

编者按

《语体文应用字汇》(以下简称《字汇》)算得上我国最早具有现代语料库意义的一项成果。它产生于上世纪20年代初，距今90余年。《字汇》常见有3个版本，原版1922年发表于《新教育》杂志第5卷第5期987页至995页。本次重刊的即是依据1922的版本，只是将当初的繁体字转为简体。含完整字频表的《字汇》曾于1928年由商务印书馆刊印发行。再有，2008年出版的《陈鹤琴全集》(第6卷)收录了该《字汇》。

《字汇》的主要贡献有：1）我国最早的汉语语料库建设及应用；2）陈鹤琴指出汉字的排序与相对频率成反比，这一点在1935年才由George Zipf提出，即广为流传的“齐夫定律”；3)《字汇》的编写主要服务于基础阶段汉语教学。也确曾作为陶行知、朱经农编写《平民千字课》的用字依据；4）从教学法的角度来说，陈鹤琴还明确提出优先教授常用字词，罕用词次之的教学理念。

值此《语料库语言学》创编之际，特重刊《语体文应用字汇》，以彰显陈鹤琴先生在我国语料库语言学研究中所作出的开创性贡献。

为尊重历史，对于《字汇》中的词语、数字和标点的使用，我们都原文照录。与现今汉语不一致之处，请勿以语病视之。

此次《字汇》文稿誊录和校对得到陈晨、陈浪、房印杰、黄俊、刘守智、王波、吴边、吴进善、章宗婧等的协助，在此一并致谢。

I. 语体文应用字汇之重要

（1）可用为小学校基本字汇

近来各地小学校逐渐采用语体文了，但采用语体文的第一个问题，就是规定语体文应用字汇，因为选择语体文教材，应当先知道那一种的字是最通用的，最通用的字共有几多，又必须知道那一种字是儿童应该先学的，那一种字应该后学的。这些问题非常要紧。解决这些问题最简单而且最有效力的方法，即在借重语体文应用字汇。

（2）可用为编制测验的根据

现在美国常用教育测验来改进教学法，分别年级等，如小学默读测验，默字测验，可以知道学生的默读与默写能力，是由测验所得的结果，可把那些最好的学生给他升级，最差的学生给他留级，或施行特别教法以图补救。这样一方可以知道学生的学力，一方可以知道教授的效力；两点晓得，在改进学生的学力上和改良教员的教法上就都有根据了。不

过这种测验的编制，必须根据确实且切实用，方有效果。语体文应用字汇，在在皆确实调查，且系最近的调查。既然可靠，又切实用。用来作为编制测验的根据，虽不敢说丝毫无弊，却比较凭意编制或根据其他字汇编制的测验，必定可靠多多了。

（3）可用为成人教育之工具

成人教育在中国成为一个大问题，因中国大多数的成人都是不识字的。倘使我们要解决这个问题，

（987页）

必须创办成人学校或各种教育成人的机关。但成人求学的时期很短，而且所学必须简易切于实用。费时要少，收效要大的法子虽然不少，采用语体文的教授，却是快捷方式之一。用此方法，语体文字汇就又占着重要的地位了。

II. 字汇研究之历史

英文应用字汇有几种：爱耳司（Ayres）曾用书信报章文学小说，以及学生的著作，共得字数368000，从中摘取单字1000，作为英文基本字汇。江司（Jones）搜集自二年级到八年级学生1050人的著作75000篇，制成字汇，共得字数15000000，单字个数4532。爱笃生（Anderson）搜集3723篇成人所做的文字，代表35种职业，共得字数361184，单字个数9223。最近桑戴克（Thorndike）费了许多年的功夫，也做了一本字汇，共同字数4565000，单字个数10000，所研究的材料共四十一种。以上所举特就几种重要的言之，至于那些不十分重要的研究略而不说了。

中文应用字汇也有几种：Pastor P. Kranz根据Soothhill的研究，编造常用四千字录。此外尚有欧美教士所做的字汇研究，因没有详细调查，无须报告。

III. 字汇研究之方法

中国字汇的分部非常复杂，编造测验初步的手续，若采用旧字汇分部方法，归纳起来，麻烦非常。为便利计，用“永”字八法作为字汇编造初步分部的标准。凡以“点”开始的字都归入“点”部，以“横”开始的字都归入“横”部。如“天”字归入“横”部，“江”字

（988页）

归入“点”部是，余类推。但中国字用“永”字八法来分部，属于“钩”部的字简直没有，属于“捺”部的字也很罕见。分部归纳的时候，把所搜集材料所有的字分别归纳于适宜各部。同样的字用符号记数法合计起来，计数的符号，或用“正”字笔画或用“册”记号笔画，“册”记号比“正”字更觉便利。归纳后，所有的单字，依用的次数之多寡，排列先后。本字汇的研究两年前即着手进行，现在已有初步的结束，大约一年后即可完竣。本字汇研究方法，计有两种：第一，专研究个别的单字；第二，研究联词和有独立的意义的单字。比方“今天早晨我进学校去读书”一句话，用第一种研究方法，

就是把这句话的字逐个归入各部，若用第二种研究法，那么“今天”“早晨”“学校”“上课”等各为一联词，照第一字开始一画，依第一种分部法归入各部。“我”“进”“去”等字都是有独立意义的单字，也要把他归入相当部内。凡专名词一概除去不计。论其用途，第一种研究效用较小，不过藉此可以知道中国语体文通用单字共有几多，通用单字应用次数之多寡，与其价值之轻重。第二种研究效用浩大，他的材料，与结果容完竣后一同报告。

IV. 研究的材料与结果

所搜集的材料都是语体文，共分六大类：（1）儿童用书，（2）新闻报，（3）杂志，（4）小学生课外著作，（5）古今小说，（6）杂类。共计554498个字，4261个单子。在这些字之中，有的用的次数很多，有的用的次数很少，现在把字汇材料，用字次数，与字汇举例三表

（989页）

A 字汇材料表

同列于下：

儿童用书类

书　　名	卷　　册	字之数目
全世界的小孩子（中）	第一集	4402
同上	第三集 第四集	9489
同上	第五集 第六集	9135
儿童文学故事（中）	第一集	811
同上	第二集	1241
同上	第三集	1433
儿童文学小说（中）	第一集	1030
儿童诗歌（商）	第一册	1267
同上	第二册	1055
儿童丛书笑话（中）	第一集	1420
同上	第二集	1097
儿童世界（商）	第一卷 第一期	4803
同上	第二卷 第八期	8175

(待续)

（续表）

书　　名	卷　　册	字 之 数 目
童话喜鹊林（商）	第一集 第九编	8548
童话无猫国	第一集 第一编	559
童话骡大哥	第一集 第七九编	1527
童话牧羊郎官	第一集 第八六编	2318
童话兔娶妇	第一集 第八一编	1200
童话书呆子	第一集 第八三编	3064
小朋友（中）	第一期 第二二期	10571

（990 页）

同上	第一四期	2229
同上	第一五期	1022
同上	第二〇期	1477
故事读本（商）	甲编　　乙编	14463
儿童世界游记（商）	第一册　第二册	10469
常识谈话（商）	第二　第八　第九	11444
同上	第四　第七	7530
同上	第一	2665
同上	第三	2849
	以上共计	127294

报刊类

报刊名	号　　期	字 之 数 目
厦门通俗教育报		3352
黑龙江通俗教育报		6913
努力周报		3696
时事新报：工商界，合作，经济		47915
同上		9188

（待续）

(续表)

报刊名	号　　期	字之数目
山东通俗白话日报		8276
北京半周刊	自十五号至二十号	26640
注音字母报	第七四期　第八六期	10048/10046
同上	第七九期　第八〇期	27393
同上	第一零九期　第一一〇期	
同上	第八五期　第七二期　第九一期	9928
	以上共计	153344

（991页）

杂志类

杂　志　名	卷　　号	字之数目
妇女杂志	第七卷 第三号	38007
同上	第七卷 第七号	52135
	以上共计	90142

小学生课外著作类

种　　名		字之数目
南高附小学生新闻		21002
同上		18520
同上		12285
	以上共计	51807

古今小说类

小说名	卷　　回	字之数目
红楼梦	第八五 九六 一〇八 一一三回	24068
水浒演义	第二 一四 二八 五六 一一二 四〇回	9641
西游记	第七 二六 四一 七四 九九回	15171

(待续)

（续表）

小 说 名	卷　　　回	字之数目
北京燕三小姐	全一册	20399
礼拜六	第一七三期	1990
	以上共计	71267

（992页）

杂类

种　　名	卷　　册	字之数目
学生婚姻问题	一册	71780
劝种小麦浅说	一册	4651
官话国耻演说	一册	23640
圣经	选新旧两经	10453
	以上共计	60625
	总计554498	

B　用字次数表

次数	字数	次数	字数	次数	字数	次数	字数	次数	字数
1	574	26	27	51	9	76	9	101–150	194
2	336	27	29	52	12	77	6	150–200	117
3	205	28	30	53	16	78	8	201–250	85
4	174	29	30	54	7	79	4	251–300	75
5	128	30	24	55	9	80	9	301–350	46
6	108	31	31	56	14	81	12	351–400	37
7	106	32	21	57	8	82	12	401–450	28
8	82	33	16	58	12	83	10	451–500	31
9	74	34	16	59	20	84	6	501–600	36
10	77	35	13	60	6	85	12	601–700	29
11	58	36	18	61	10	86	5	701–800	28

（待续）

(续表)

次数	字数	次数	字数	次数	字数	次数	字数	次数	字数
12	65	37	10	62	8	87	4	801–900	15
13	62	38	17	63	8	88	4	901–1000	15
14	49	39	24	64	11	89	6	1001–2000	66
15	46	40	16	65	10	90	4	2001–3000	19
16	45	41	23	66	8	91	5	3001–4000	5
17	45	42	30	67	15	92	7	1001–5000	4
18	51	43	20	68	11	93	10	5001 以上	10
19	26	44	12	69	12	94	7		
20	38	45	12	70	8	95	6		
21	40	46	15	71	11	96	3		
22	27	47	15	72	6	97	5		
23	28	48	14	73	10	98	5		
24	33	49	10	74	9	99	6		
25	26	50	10	75	7	100	7		

（993 页）

C　语体文应用字汇举例

从上面用字次数表看来，次数愈小，字数愈大；次数愈大，字数愈小。次数与字数适成反比例，所谓次数就是指在以上554498字中，所遇见的次数也。凡次数很少的字，他的用处和价值可以说是很小。次数很大的字，他的用处和价值可以说是很大。如次数“1”的574字用处极小，而次数“5000”的10字用处极大是。

次数71–80	次数41–45	次数21–25	次数9	次数5	次数1
71郎	41趋	21舰	仓	弦	喀
73革	42覆	22蟹	馒	辉	姣
75菜	43源	23伏	型	霹	崎
78席	44璃	24盼	痴	陨	徕
80制	45稳	25箱	豹	礁	讹

(待续)

（续表）

次数81–90 81巨 83施 85顿 87钟 90灭	次数46–50 46撤 47睛 48赖 49贯 50逼	次数26–30 26腐 27耐 28藉 29晒 30鸽	次数10 矫 庞 宪 贡 迈	次数6 刦 劾 迴 驮 腥	次数2 缄 肪 绦 麟 敖
次数91–100 91泪 93桌 96摇 98庄 100喝	次数51–60 51梧 53纯 55剧 58需 60扫	次数31–35 31逸 32畜 33牌 34椅 35垫	次数11–15 11喧 12堕 13妾 14宴 15捞	次数7 擎 署 肄 舐 舂	次数3 轫 颂 酪 贸 翔
次数101–150 101防 112顶 121响 140惊 148错	次数61–70 61薄 62迷 63添 66整 70谷	次数36–40 36偶 37硬 38箭 39粗 40污	次数16–20 16埋 17宾 18注 19浊 20蓝	次数8 环 穗 熊 寝 冈	次数4 簿 肃 蝎 垓 帆

（994页）

用字汇举例

次数 151–200	次数 351–400	次数 701–800	次数 2001–3000
式151 致164 奏172 场184 玩200	船351 拿365 怕378 造387 钱400	主701 其725 常768 相787 未796	么2001 里2361 得2655 你2882 去2989
次数 201–250	次数 401–450	次数 801–900	次数 3001–4000
料201 狗211 渐232 换241 观250	运404 觉428 姐436 山440 父447	些807 劳825 定848 各856 法878	道3108 也3035 上3338 子3528 说3626

（待续）

(续表)

次数 251–300	次数 451–500	次数 901–1000	次数 4001–5000
哥 251 记 264 买 272 倒 280 商 300	甚 451 着 460 伐 479 高 485 路 498	当 908 情 921 从 962 呢 977 教 996	可 4358 这 4427 来 4507 在 4533
次数 301–350	次数 601–700	次数 1001–2000	次数 5001 以上
尽 301 权 326 产 330 电 343 念 350	向 601 现 620 则 640 故 680 月 698	经 1013 年 1203 学 1585 见 1872 生 1992	们 5090 有 6453 我 7955 不 8953 的 20228

志谢：以上所报告的研究，非有敝校教育科之补助，诸同事之指导，并李君尚春以及其他各助理细心合作，断难成全，因此特附笔致谢。

（995 页）

附注

《字汇》1922、1928、2008 三个版本的电子文档可由 http://www.bfsu-corpus.org/static/corpus-classics/ 下载得到。

陈鹤琴简历

陈鹤琴（1892–1982），浙江上虞人，我国著名儿童教育家、儿童心理学家。早年毕业于清华大学，留学美国五年，1919 年获得哥伦比亚大学硕士学位。陈鹤琴回国后，最初任南京高等师范学校教授，国立东南大学（后更名国立中央大学、南京大学）成立后，任教授兼教务主任。在此期间，他致力于研究儿童心理学和幼儿教育学。1922 年发表《语体文应用字汇》，为第一项基于大规模真实语料的汉字字频研究成果，开创了我国汉字字量的科学研究。对编写小学课本和普及教育起了推动作用。

会讯动态

第36届ICAME大会征文通知

第36届ICAME大会将于2015年5月27号至31号在德国特里尔召开。

本次大会的主题是：Words, words, words—Corpora and lexis。

会议将邀请Kate Burridge、Thomas Herbst、Graeme Trousdale、Edmund Weiner作主旨发言。

大会提交摘要的截止日期是2014年12月1日；录用通知寄发时间是2015年1月15日。

会议的官方邮箱是：icame36@uni-trier.de

会议的官方网址是：http://www.uni-trier.de/index.php?id=52257

“国际学习者语料库研究会”成立

“国际学习者语料库研究会”于2013年9月在“第二届学习者语料库研讨会”于挪威卑尔根正式成立。研究会将定期举办双年会。

国际学习者语料库研究会旨在推动各国学习者语料库的创建，支持新的研究方法和分析工具的创新。研究会还特别主张学习者语料库研究与语言习得理论、普通语言学理论、语言教学、语言测试、自然语言处理领域的结合。

协会的创始会员是比利时鲁汶天主教大学英语语料库研究中心的Gaëtanelle Gilquin、Sylviane Granger、Fanny Meunier和Magali Paquot。

研究会的官方网站是：http://www.learnercorpusassociation.org

“ToRCH2009现代汉语语料库”建成

ToRCH2009语料库是按布朗语料库（Brown corpus）取样方案创建的100万词次现代汉语语料库。该语料库所收15种文本类别情况，可见下表。

文类代码	体裁类型	文类代码	体裁类型
A	新闻报道	J	学术、科技
B	社论	K	普通小说
C	报刊评论	L	侦探小说
D	宗教	M	科幻小说
E	日常技艺及消遣爱好	N	历险悬疑小说
F	通俗读物	P	言情小说
G	传记、回忆录等	R	喜剧幽默
H	政府或机构公文及文宣		

ToRCH2009语料库创建的突出特点是“共建共享”。它是由全国64所高校的115位老师和硕博士生参与语料收集，共同创建的现代汉语语料库。ToRCH2009语料库项目由北京外国语大学中国外语教育研究中心许家金教授发起，并统筹语料收集、整理和校对工作。

该语料库大小为1,066,347词（计1,670,356字）。ToRCH2009中所收文本绝大部分出版于2009年。语料库名称ToRCH为Texts Of Recent CHinese的缩略词。我们希望ToRCH语料库能以类似的取样模式，每隔若干年新建一库，比如ToRCH2014、ToRCH2019、ToRCH2024，从而可以用于考察现代汉语的动态发展。因此，我们希望ToRCH语料库将来能形成系列。此次的ToRCH2009是该系列的第一个语料库。语料库取名ToRCH也表达了希望该语料库系列可以“薪火相传”，不断延续的含义。

该语料库与此前创建的Crown和CLOB语料库（参看：http://icame.uib.no/ij37/Pages_175-184.pdf）构成汉英可比语料库（comparable corpora），可用于英汉对比研究。

Crown、CLOB及ToRCH2009语料库皆可通过BFSU CQPweb语料库平台（http://124.193.83.252/cqp/）在线检索。

ToRCH2009语料库的文本可从http://www.bfsu-corpus.org/channels/corpus下载。

《语料库语言学》稿约

一、本刊欢迎以下各类来稿：

1. 语料库语言学理论探索
2. 语料库与中介语研究
3. 语料库与语言对比研究
4. 语料库与翻译研究
5. 语料库与语言描写
6. 语料库与话语研究
7. 语料库研究新方法
8. 语料库软件的设计与开发
9. 语料库的研制与创建
10. 书刊评介（语料库相关书籍的述评。所评介的书籍限近3年出版的高水平论著）
11. 以上未能涵盖的其他相关研究。

二、请将研究性论文稿件字数控制在8,000–10,000字以内（字数计算含中文标题、中文摘要、正文、参考文献）。书评字数5,000字上下为宜。投稿稿件中请另附英文标题及英文摘要。

三、初审、复审在编辑部内完成。初审由编辑部指定的本单位编辑或专家，复审由主编完成。本刊按刊物实际载文量200%的比例，筛选稿件寄送一至两名国内外同行专家匿名外审。

四、本刊实行同行专家匿名审稿制。一般审稿周期为3个月。

五、不收取任何审稿费、版面费，亦不支付稿费。被录用者赠送当期杂志4册。

六、投稿方式：本刊不接受纸质投稿。请将稿件电子版Word文档发送至bfsucrg@sina.com。

本刊格式体例

（同《外语教学与研究》杂志格式）

1. 稿件构成

论文中文标题、中文提要、中文关键词、论文正文（含参考文献）

论文英文标题、英文提要（另页）

附录等（若有）

作者姓名、单位（中英文）、通讯地址、电话号码、Email地址（另页）

2. 提要与关键词

论文须附中、英文提要；中文提要200–300字，英文提要150–200词。另请给出能反映全文主要内容的关键词3–5个。

3. 正文

3.1 结构层次

正文分为若干节，每节可分为若干小节。

3.2 标题

节标题、小节标题独占一行，顶左页边起头。

节号的形式为1、2、3……，节号加小数点，然后是节标题；或一、二、三……，节号后加顿号，然后是节标题。

小节号为阿拉伯数字，形式为1.1、1.2、1.3……，1.1.1、1.1.2、1.1.3……。小节号后空1格，不加顿号或小数点，然后是小节标题。

小节之下可以采用字母A. B. C.，a. b. c.，(a) (b) (c)或罗马数字I. II. III.，i. ii. iii.，(i) (ii) (iii) 对需要编号的内容加以编号。

3.3 字体

正文的默认字体为宋体五号。

中文楷体用于字词作为字词本身使用，如：

“楷字怎么念？”

英文倾斜字体的使用范围主要是：

（1）词作为词本身使用，如：

The most frequently used word in English is *the*.

（2）拼写尚未被普遍接受的外来词，如：

Jiaozi is very popular in China.

（3）书刊等的名称。

3.4　图表

图标题置于图的下方，表标题置于表的上方。

图号/表号的格式为“图/表+带小数点的阿拉伯数字”。

3.5　参引

一切直接或间接引文以及论文所依据的文献均须通过随文圆括号参引标明其出处。

参引的内容和语言须与正文之后所列参考文献的内容和语言一致。

作者名字若是英文或汉语拼音，不论该名字是本名还是译名，参引时都仅引其姓。其他民族的名字或其译名若类似英文名字，参引时比照英文名字。

转述某作者或某文献的基本或主题观点或仅提及该作者或该文献，只需给出文献的出版年，如：

陈前瑞（2003）认为，汉语的基本情状体分为四类，即状态、活动、结束、成就。

直接或间接引述某一具体观点，须给出文献的页码，格式是“出版年：页码”，如：

吕叔湘（2002：117）认为，成做动词时，有四个义项：1）成功、完成；2）成为；3）可以、行；4）能干。

如作者的名字不是正文语句的一个成分，可将之连同出版年、页码一起置于圆括号内，如：

这是社交语用迁移的影响，即“外语学习者在使用目的语时套用母语文化中的语用规则及语用参数的判断”（何兆熊 2000：265）。

圆括号内的参引若不止一条，一般按照出版年排序。同一作者的两条参引之间用逗号隔开，如：Dahl（1985, 2000a, 2000b）；不同作者的参引之间用分号隔开。

文献作者若是两个人，参引时引两个人的名字。中文的格式是在两个名字之间加顿号，如“吕叔湘、朱德熙（1952）”；英文的格式是在两个姓之间加&号，表示‘和’，如Li & Thompson（1981）。

文献作者若是三人或三人以上，参引时仅引第一作者的名字。中文的格式是在第1作者名字之后加“等”字，如“夸克等（1985/1989）”；英文的格式是在第一作者的姓之后加拉丁缩略语“*et al.*”，如“Quirk *et al.*（1985）”，*et al.* 为斜体。

3.6 随文圆括号夹注

除了用于参引外，随文圆括号夹注主要用于提供非常简短的说明、译文的原文以及全名的缩写或全称的简称，如：

对于莎士比亚研究学者来说，最重要的词典有两部：一部是19世纪70年代德国人Alexander Schmidt以德意志民族特有的勤奋及钻研精神编纂的两卷本巨著*Shakespeare Lexicon and Quotation Dictionary*（1874/1902/1971，以下简称Lexicon），另一部是*Oxford English Dictionary*（1884-1928/1989，通常简称OED）。

随文夹注的字体与正文默认字体相同。

3.7 附注

一般注释采用附注的形式，即在正文需注释处的右上方按顺序加注数码1、2、3……，在正文之后写明“附注”或“注释”字样，然后依次写出对应数码1、2、3……和注文，回行时与上一行注文对齐。

3.8 例证/例句

例证/例句宜按顺序用（1）（2）（3）……将之编号。每例另起1行，左缩进1个中文字符。编号与例句之间不空格，回行时与上一行例证/例句文字对齐。外文例证/例句可酌情在圆括号内给出中译文。

4. 参考文献

每一条目首行顶左页边起头，自第2行起悬挂缩进2字符。

文献条目按作者姓氏（中文姓氏按其汉语拼音）的字母顺序排列。

中文作者的姓名全都按姓+名的顺序给出全名。英文仅第一作者的姓名（或汉语拼音姓名）按照姓+名的顺序给出，姓与名之间加英文逗号，其他作者的姓名按其本来顺序给出。英文作者的名仅给出首字母。

中外文献分别排列，外文在前，中文在后。

同一作者不同出版年的文献按出版时间的先后顺序排列，同一年的出版物按照文献标题首词的顺序排列，在出版年后按顺序加a、b、c以示区别。

外文论文（包括学位论文）的篇名以正体书写，外文书名以斜体书写。篇名仅其首词

的首字母大写，书名的首词、尾词以及其他实词的首字母大写。

篇名和书名后加注文献类别标号，

专著标号为[M]

论文集为[C]

论文集内析出文章为[A]

期刊论文为[J]

尚未出版之会议论文为[R]

博士论文和硕士论文为[D]

词典为[Z]

网上文献为[OL]

期刊名称后的数字是期刊的卷号，通常是每年一卷，每卷统一编页码。如没有卷号只有期号，则期号须置于圆括号内；如有卷号但每期单独编页码，须在卷号后标明期号并将期号置于圆括号内。

每条顶左页边起头，回行时悬挂缩进2个中文字符。

English Abstracts

Some reflections on Corpus Linguistics upon request

.... GUI Shichun（1）

This is an interview with Professor Gui Shichun of Guangdong Foreign Studies University on his personal history of corpus linguistics research. Professor Gui is one of the forerunners of English corpus linguistics studies in China. His Chinese Learners English Corpus (in collaboration with Professor Yang Huizhong) has been the empirical basis for enormous amount of research projects on Chinese learners' interlanguage. He has kept a close eye on cutting-edge corpus technologies and methodologies. At the age of 79, he published his Biberian multidimensional study of linguistics research articles on a self-compiled corpus. More recently, he has been learning R and implementing it in English and Chinese statistical analyses. This more or less reflective account of personal corpus research will be food for thought for the young generation of corpus linguists.

Revisiting English collocational frameworks as units of meaning

....HE Anping（16）

Aiming at meaningful patterns of collocational frameworks in the perspective of corpus linguistic phraseology, this study adopts the concept of "extended framework" and the method of "cumulative frequency" to explore 3 types of collocation frameworks: "the ? of ", "a ? of " and "the ? in" in terms of lexical and grammatical co-occurring patterns. It is found that these "Article + Noun + Preposition + Noun" sequences tend to: 1) structurally, share an embedded and recurring patterning; 2) semantically, demonstrate a semantic preference in describing target nouns in each framework, thus showing grammar words on each side of a framework constrain its pattern meaning; and 3) functionally, play a role of signalling and linking propositions in a discourse, which is more salient a feature in the first two frameworks.

Corpora, the Plain Meaning Rule and litigation proofs in the US law

....LIANG Maocheng（25）

Corpora are not only used in linguistic studies, but also in other related research. A good case in

point is the popular use of corpora in U.S. courts for the purpose of interpreting legal statutes in accordance with the Plain Meaning Rule. On the other hand, corpora could also play an important role in finding and providing litigation proofs. This paper introduces a few U.S. legal cases, where corpora were used for determining the plain meaning of words in legal statutes and providing litigation proofs. It is hoped that the study could provoke ideas among Chinese researchers and practitioners in law and linguistics.

A study of the macro linguistic features of the English translations of the Four Chinese Classic Novels

....LIU Zequan & LIU Dingjia（34）

Based on the statistics from the Chinese-English parallel corpus the "Four Chinese Classic Novels" and their multiple English translations, this study makes a comparative investigation of the textual features of the 11 English translations, including their lexical diversity (in terms of the STTR and noun-pronoun ratio), lexical density, average length in word and sentence, as well as C-E sentence alignment type. The investigation is oriented at a description of the translation strategies employed by the respective translators, the similarities and differences of the translator style of each version and the reasons behind them. It is found that while the translations display quite a few similarities between them, which all pertain to the features of explicitation, each reveals its own uniqueness in so far as their respective translator style is concerned.

A study of the use of chunks in English research articles by Chinese EFL scholars

....ZHENG Honghong（47）

This study attempts to analyze the most frequently used chunks in published English research articles (RAs) written by Chinese EFL scholars by comparing them with those used in RAs written by international scholars. The most frequently used chunks in the Chinese Academic Writing English Corpus (CAWEC) and the Foreign Academic Writing English Corpus (FAWEC) were retrieved and compared in terms of their frequencies, functions and structures. Findings of the study indicate that: 1) Chinese EFL scholars tend to overuse a limited variety of chunks; 2) Chinese EFL scholars tend to use more functionally simple chunks and content-related chunks; 3) the chunks used by Chinese EFL scholars are structurally similar to those used by international scholars; and 4) Chinese EFL scholars use more chunks that are characteristic of the spoken register.

Research on semantic prosody: Theory, methodology and application

.... LU Jun（58）

This paper reviews the development of the theoretical framework, research methodology and application areas of semantic prosody. It begins with a brief description of multiple schools of thoughts on defining semantic prosody, followed by an analysis of their contribution to linguistic studies and the relationship among them. Then, research methods are reviewed from the perspectives of their theoretical basis, potential improvement and future development. The third part deals with the application areas of semantic prosody in terms of their principle, theoretical contribution and future research direction. Finally, it is concluded that a basic theoretical framework and research methodology for semantic prosody research have been established, which has a strong demand on further extended research.

An overview of corpus linguistics research in China (1998–2013): A CiteSpace based analysis

.... LIU Xia, XU Jiajin & LIU Lei（69）

This article reviews corpus based language studies drawing on the CSSCI indexed journal articles between 1998 and 2013. The bibliographic data have been quantified and visualized with the bibliometric tool CiteSpace. The results show that corpus based language studies have been carried out along major lines of research: corpus based interlanguage analysis and bilingual corpus based translation studies and contrastive linguistics over the 16 years span. During the earlier years in the time span of the review, focus of attention was devoted to learner corpus research, while bilingual corpus based studies have become more popular in recent years. The cross-fertilization of different disciplines and research fields highlights the corpus research in last three years.

Construction of and research on parallel Russian-Chinese corpora: With special reference to academic texts of social sciences and humanities

.... TAO Yuan（78）

Parallel corpora construction is a prerequisite for descriptive translation studies. Although there are a variety of such corpora, bilingual or multilingual, both at home and abroad, there are few

parallel Russian-Chinese ones. This paper attempts to illustrate the construction of such corpora in terms of corpora structure, text selection and construction procedure. Besides sentential alignment, header design and intersection retrieval, this article also discusses in detail: 1) how to retrieve Russian words in inflectional forms in the corpora, and 2) how to generate termbase automatically. Given that corpora design and research are closely related to the research purpose, this article also discusses the future directions of research based on parallel Russian-Chinese corpora.

Characters used in vernacular Chinese

.... CHEN Heqin（94）

Characters Used in Vernacular Chinese compiled by Heqin Chen and his associates in early 1920s was a milestone in corpus based Chinese studies. A brief report about the character list was first published in the journal of *New Education* in 1922. The complete version of the list was published as a booklet in 1928. A more accessible version of the list is the reprint in an edited volume (2008: 55-114). Some major contributions of Chen's Character List to Chinese corpus linguistics and its application can be highlighted as follows: 1) The project was one of the first Chinese corpus projects in the modern sense, because it followed principled sampling considerations and collected a sizeable quantity of authentic texts; 2) In the preface to the book, Chen (1922: 994) observed the overall tendency of the inverse proportionality between relative frequency and character rank as recognized by George Zipf (1935); 3) The project was consciously motivated by pedagogical ends, and used as vocabulary control criterion in the Chinese textbook *Early Chinese Lessons for Illiterates* compiled by Tao and Zhu; and 4) Chen explicitly stated that the vocabulary teaching methodology of "the frequent than the rare" (Chen 1922: 987, 994).